55 RESPUESTAS A DUDAS TÍPICAS DE ESTADÍSTICA

ROBERTO BEHAR GUTIÉRREZ
PERE GRIMA CINTAS

DIAZ DE SANTOS

Subido por:

Libros de Ingeniería Química y más

https://www.facebook.com/pages/Interfase-IQ/146073555478947?ref=bookmarks

Si te gusta este libro y tienes la posibilidad, cómpralo para apoyar al autor.

55

Respuestas a dudas típicas de ESTADÍSTICA

55

Respuestas a dudas típicas de ESTADÍSTICA

Roberto Behar Gutiérrez
Pere Grima Cintas

DIAZ DE SANTOS

© Roberto Behar Gutiérrez, Pere Grima Cintas, 2004

Reservados todos los derechos.

«No está permitida la reproducción total o parcial de este libro,
ni su tratamiento informático, ni la transmisión de ninguna
forma o por cualquier medio, ya sea electrónico, mecánico
por fotocopia, por registro u otros métodos, sin el permiso
previo y por escrito de los titulares del Copyright.»

Ediciones Díaz de Santos, S. A.
Doña Juana I de Castilla, 22. 28027 Madrid
España

Internet: http://www.diazdesantos.es/ediciones
E-Mail: ediciones@diazdesantos.es

ISBN: 84-7978-643-4
Depósito legal: M. 33.434-2004

Diseño de cubierta: A. Calvete
Fotocomposición: Fer
Impresión: Edigrafos
Encuadernación: Rústica-Hilo

Impreso en España

ACERCA DE LOS AUTORES

Roberto Behar Gutiérrez es Licenciado en Educación en la especialidad de Matemáticas, por la Universidad Santiago de Cali, y es Estadístico por la Universidad del Valle (Cali, Colombia). Obtuvo el grado de doctor en la Universidad Politécnica de Cataluña. Es profesor de la carrera de estadística de la Universidad del Valle desde su fundación en 1978, donde también ha sido director del departamento de Producción e Investigación de Operaciones, director de la carrera de estadística, y director del Master en Ingeniería Industrial y de Sistemas.

Ha sido asesor estadístico para diversas instituciones colombianas en estudios sobre medio ambiente y desarrollo social. También ha asesorado a empresas sobre temas relacionados con el control estadístico de la calidad y estadística industrial. Entre sus publicaciones se encuentra el libro de texto que escribió junto a Mario Yepes: "Estadística: Un enfoque descriptivo". Ed. Feriva, 1995. Ha escrito también numerosos artículos sobre técnicas estadísticas en revistas especializadas, algunos de los cuales tratan sobre su enseñanza, como el publicado con Mario Miguel Ojeda en el *Newsletter* (1997) del *International Statististical Institute. (ISI): "A Reformulation of the problem of Statistical Education: A learning Perspective*" o el que escribió con Pere Grima en la revista "Estadística Española" (2001): "Mil y una dimensiones del aprendizaje de la estadística". También ha realizado numerosas conferencias sobre el uso de la estadística y sobre su enseñanza y aprendizaje, tema que es una de sus pasiones.

Pere Grima Cintas es doctor ingeniero industrial y profesor de la Universidad Politécnica de Cataluña, donde también es coordinador académico del Master en Gestión de la Calidad. Su especialidad son las técnicas estadísticas para el control y la mejora de la calidad, tema sobre el que ha asesorado a numerosas empresas e instituciones. También se ha ocupado de temas relacionados con la gestión de la calidad y en el año 2000 fue evaluador de la *European Foundation for Quality Management* (EFQM) para el premio europeo a la calidad que otorga esta institución. En el periodo en que ha estado trabajando en este libro, sus actividades de asesoramiento a empresas se han centrado mayoritariamente en la implantación de programas de mejora Seis Sigma.

Junto con sus compañeros Albert Prat, Xavier Tort-Martorell y Lourdes Pozueta escribió el libro "Métodos estadísticos. Control y mejora de la Calidad", Ediciones UPC, del que ya se han realizado varias ediciones y que ha sido publicado en Iberoamérica por Editorial Alfaomega. Con Xavier Tort-Martorell escribió "Técnicas para la Gestión de la Calidad", editado por Díaz de Santos en 1995. También le gusta dedicar parte de su tiempo a trabajar en temas relacionados con la divulgación de la estadística.

Presentación

Muchos de los que alguna vez hemos sido estudiantes de un curso de Estadística, recordamos momentos en los que intentábamos entender, no siempre con éxito, las razones por las cuales había que hacer las cosas de una determinada manera. ¿Por qué dividir por n-1 al calcular la desviación estándar? ¿Por qué no dividir por el número total de datos? El tiempo disponible en los cursos no permite explicarlo todo y en ocasiones el profesor, en su intento por dar una explicación al estudiante que pregunta, responde usando términos como "grados de libertad" o "estimador insesgado", lo cual puede generar más dudas de las que aclara.

Por otra parte, a través de nuestra experiencia ayudando a profesionales de la Medicina, la Administración o la Ingeniería en el uso de métodos estadísticos, hemos comprobado que los conceptos o las técnicas con que se trabaja no siempre están del todo claras. ¿Cómo hay que interpretar el p-valor que da el listado del ordenador? ¿Es lo mismo diferencia significativa que diferencia importante? Esta falta de seguridad en su manejo hace que muchas veces se evite hacer uso de todas las posibilidades que brinda la estadística, con lo que se pierde la oportunidad de obtener una información que puede resultar muy útil para la toma de decisiones.

También en un ámbito no estrictamente profesional existen muchas dudas "populares" en torno a la Estadística: ¿cómo es que con una muestra de 2.000 personas puede conocerse razonablemente bien la opinión de un país de 40 millones de habitantes?, o lo que es todavía más sorprendente, ¿cómo es que esas 2.000 personas también serían suficientes para una población de 100 millones? Y ligada con estas, si la Estadística es tan potente, ¿por qué cuesta tanto acertar en los sondeos electorales?

Este texto pretende dar respuesta a muchas de estas preguntas y nuestra intención es que sea útil tanto a los estudiantes de los cursos de estadística que se imparten en la universidad, como a los profesionales que están interesados en refrescar sus ideas o aclarar dudas concretas, y también a todas aquellas personas interesadas en esta disciplina que quieran resolver algunas de sus dudas.

Sin ser exhaustivos, pues siempre es posible aumentar la lista con nuevas preguntas, hemos tratado de cubrir un amplio espectro, tratando dudas en estadística descriptiva, distribuciones de probabilidad, estimación, contraste de hipótesis, comparación de poblaciones, correlación y regresión, diseño de experimentos, estudios de capacidad y control de procesos y un apartado para dudas varias, como las relacionadas con los grados de libertad y el teorema central del límite, entre otras.

Muchas preguntas tienen un carácter general e introductorio y son "aptas para todos los públicos", pero otras tratan sobre temas específicos en el contexto de las ecuaciones de regresión, el diseño de experimentos o el control estadístico de procesos. En este último caso se requiere un cierto nivel de conocimientos sobre el tema, aunque si la pregunta despierta interés, seguramente ya se sabe lo suficiente para entender la respuesta. En todos los casos se ha intentado usar un lenguaje coloquial, recurriendo a la intuición y apoyándose en la metáfora, pero procurando que no haya pérdida en el rigor.

Se ha intentado también que cada respuesta sea lo más autocontenida posible, es decir, lo suficientemente completa, para que no requiera de otras para su adecuada comprensión. De todas maneras, en cada uno de los temas que se tratan, se han colocado las dudas y sus respuestas en el orden que consideramos más efectivo, de tal manera que un lector que desee leer todas las preguntas de un apartado en forma secuencial vaya ganando elementos para comprender mejor la siguiente.

Dejando claro que cualquier falta en la virtud de este trabajo es de exclusiva responsabilidad de los autores, deseamos poner de manifiesto nuestro agradecimiento a todos nuestros compañeros en las tareas docentes, seguramente la mejor fuente de información que hemos tenido. Lluís Marco, de la Universitat Politècnica de Catalunya y Guillermo de León, de la Universidad Veracruzana nos sugirieron algunas de las preguntas que se incorporan y también ideas y posibles enfoques para muchas respuestas, además de leer los originales y sugerir numerosas mejoras. Rafael Antonio Klinger y Eloina Mesa, de la Universidad del Valle, también leyeron los originales y realizaron muchas sugerencias que han mejorado notablemente la claridad de las respuestas.

Deseamos agradecer también a la Agencia Española de Cooperación Internacional (AECI) y a nuestras Universidades, la Universidad del Valle y la Universitat Politècnica de Catalunya, las ayudas y facilidades obtenidas para la realización de este trabajo.

Muy probablemente, no podremos evitar la frustración de algunos de nuestros lectores al buscar en vano alguna duda que no fue tratada aquí, o al no quedar del todo satisfechos con alguna respuesta. Nuestra aspiración es poder recoger todas las sugerencias y apreciaciones que nos permitan realizar un proceso mejora continua de nuestro trabajo, por lo que agradeceremos todos los comentarios y sugerencias que nos hagan llegar a través de la página web: www.55RespuestasEstadistica.com.

Barcelona y Santiago de Cali, Mayo de 2004

Índice

Acerca de los autores VII

Presentación IX

Estadística descriptiva

1. ¿Para qué sirve la mediana, si ya tenemos la media aritmética? 3
2. ¿Tiene alguna aplicación práctica la media geométrica? 5
3. ¿Por qué en la expresión de la varianza se utiliza el cuadrado de las diferencias en vez de su valor absoluto? 7
4. ¿Por qué cuando se calcula la varianza de una muestra se divide por n-1 en vez de dividir por n? 11
5. ¿Cuál es la forma "correcta" de calcular los cuartiles? 15
6. ¿En cuántos intervalos conviene dividir los datos para construir un histograma? ¿Qué otros aspectos hay que tener en cuenta? 17
7. ¿Cuándo conviene utilizar *boxplots* para analizar o describir datos? 21
8. En los *boxplots* las anomalías se marcan a partir de ± 1,5 veces el rango intercuartílico (IQR) ¿De dónde sale el 1,5? 23
9. ¿Qué hay que hacer cuando nos encontramos con valores atípicos? 25
10. ¿Qué es la curtosis (o *kurtosis*) y para qué sirve? 29

Distribuciones de probabilidad

11. ¿Cómo se sabe que una variable aleatoria concreta sigue una determinada distribución de probabilidad? 35
12. La media de una muestra es un número concreto. ¿Por qué se dice entonces que es una variable aleatoria? 39
13. ¿Por qué la función densidad de probabilidad de la distribución Normal es la que es? 41
14. ¿Por qué las probabilidades calculadas a través de la Normal estandarizada coinciden con las buscadas en la distribución de interés? 47
15. Yo mido 1,68. ¿Por qué la probabilidad de que una estatura sea 1,68 calculada con la distribución Normal es 0? 51
16. ¿Existen variables aleatorias que presenten un comportamiento "contrario" a la distribución Normal, siendo los valores más probables los de los extremos? 53
17. ¿De dónde sale la fórmula de la distribución de Poisson? 57
18. ¿Cómo se puede ver que la distribución de la varianza muestral está relacionada con la distribución chi-cuadrado? 59

19. ¿Por qué da un resultado distinto sumar *k* variables aleatorias de la misma distribución de probabilidad que tomar una y multiplicarla por *k*? 61

Estimación

20. Sabemos que las características de una muestra (proporción, media, ...) varían de una muestra a otra. ¿Por qué entonces creer en los resultados de una muestra, sabiendo que si tomáramos otra esos resultados serían distintos? 65

21. ¿Qué significa la expresión: "un intervalo de confianza del 95% es 27,5% ± 3,6%"? 67

22. ¿Por qué para estimar la media de una población el tamaño de la muestra no crece proporcionalmente con el tamaño de la población? 69

23. ¿Por qué cuesta acertar en los sondeos electorales? 73

24. ¿Qué es un estimador de máxima verosimilitud? 77

Contraste de hipótesis

25. ¿Qué es el p-valor y cuál es el significado de las otras palabras clave que aparecen en el contraste de hipótesis? 81

26. ¿A partir de qué p-valor es razonable rechazar la hipótesis nula? 85

27. ¿Qué tipos de error se pueden cometer en un contraste de hipótesis? 87

28. ¿Es correcto multiplicar por 2 el área de cola en los tests de igualdad de varianzas cuando H_1 es del tipo "distinto de"? 91

29. ¿Por qué respecto a la hipótesis nula se habla de "no rechazo" y no de "aceptación"? 93

30. ¿Es lo mismo diferencia significativa que diferencia importante? 95

Comparación de tratamientos

31. ¿Cómo elegir la hipótesis alternativa que conviene plantear? 99

32. Si la hipótesis alternativa es del tipo "mayor que" o "menor que", ¿cómo se sabe hacia qué lado hay que mirar el área de cola? 101

33. ¿Por qué el análisis de la varianza se llama así, cuando en realidad se trata de una técnica para comparar medias y no varianzas? 103

34. ¿Por qué para comparar *k* tratamientos se utiliza la técnica de análisis de la varianza, en vez del ya conocido test de la *t* de Student, aplicándolo a todas las parejas que se pueden formar con *k* tratamientos? 105

Correlación y Regresión

35. ¿Por qué cuando se ajusta una nube de puntos a una ecuación de regresión se utiliza siempre el criterio de minimizar la suma de los cuadrados de los residuos, y no otros como minimizar la suma de su valor absoluto? 109

36. Si los coeficientes de una ecuación de regresión son unos números concretos, ¿por qué se dice que son variables aleatorias? 111

37. ¿Por qué cuando se ajusta una recta que pasa por el origen no se utiliza el coeficiente de determinación R^2 como medida de calidad del ajuste? 115

38. ¿Por qué cuando se comparan ecuaciones de regresión con distinto número de variables regresoras no se utiliza R^2 sino el llamado R^2 ajustado? 119

39. ¿Cómo se pueden utilizar e interpretar variables cualitativas en una ecuación de regresión? 125

40. ¿Por qué del conjunto de variables candidatas a entrar en un modelo de regresión no necesariamente se seleccionan las que están más correlacionadas con la variable dependiente *Y*? 131

Diseño de experimentos

41. ¿Por qué no es una buena estrategia ir moviendo las variables una a una cuando se trata de estudiar experimentalmente cómo estas afectan a una respuesta? 135

42. ¿Cómo es posible estudiar por separado el efecto de cada una de las variables que afectan a una respuesta si, tal y como se hace en los diseños factoriales, se mueven todas a la vez? 137

43. ¿Por qué funciona el algoritmo de Yates? 143

44. ¿Por qué cuando se representan valores en papel probabilístico normal (ppn), en la fórmula que da la ordenada se resta 0,5 del número de orden? 147

45. En los diseños factoriales, ¿cómo se puede escribir una ecuación para la respuesta a partir de los efectos? 151

46. ¿Qué es un diseño bloqueado? ¿Por qué en estos diseños no se tienen en cuenta las interacciones entre los factores de bloqueo y el resto de factores? ¿Qué ocurre si esas interacciones existen? 155

47. ¿Por qué es razonable suponer no significativas las interacciones de 3 o más factores? 159

48. ¿Qué hacer si al aleatorizar el orden de experimentación se obtiene el orden estándar de la matriz de diseño? 163

Estudios de capacidad y control estadístico de procesos

49. ¿Qué diferencia hay entre un estudio de capacidad a corto y largo plazo? ¿Cómo se estima la variabilidad en uno y otro caso? 167

50. ¿Por qué en los gráficos de control es más eficiente controlar medias que observaciones individuales? 171

51. En los gráficos de control, ¿la línea central debe ser el valor objetivo o el promedio obtenido al hacer el estudio de capacidad? 175

Varios

52. Cuando se habla de transformación logarítmica, ¿se refiere al logaritmo decimal o al neperiano? 179

53. ¿Qué significan los llamados “grados de libertad”? 181

54. ¿Debe decirse “Teorema central del límite” o “Teorema del límite central”? 185
55. ¿Cuál es la mejor estrategia para ganar la lotería (nacional, primitiva,...)? 187

Créditos y referencias

¿Cómo hemos resuelto nuestras dudas? 193
Libros y páginas web que se citan 197

Estadística descriptiva

1 ¿Para que sirve la mediana, si ya tenemos la media aritmética?

La media aritmética es una excelente medida de tendencia central. Sus buenas propiedades, junto con el hecho de ser fácil de entender y de calcular, la hacen muy usada y también muy apreciada (a veces demasiado, como cuando se pretende resumir solo en ella toda la información que contienen los datos), pero la mediana tiene unas propiedades de las que carece la media, por lo que es un buen complemento informativo e incluso en algunos casos puede ser una medida más útil. Estas propiedades son:

- Es más robusta que la media frente a la presencia de anomalías. Supongamos que nuestros datos son: 2, 5, 6, 7 y 9. La media es 5,6 y la mediana es 6. Si al introducir los datos al ordenador nos equivocamos y en último lugar en vez de 9 introducimos 99, la media pasa a ser de 23,8 mientras que la mediana sigue siendo 6.

  En algunos casos, como cuando se trabaja con datos todavía no depurados, fijarse en la mediana puede ser más recomendable porque la información que da está menos afectada por las posibles anomalías que puedan existir.

- Por su propia definición, la mediana deja un 50% de las observaciones por encima y otro 50% por debajo y esto le da unas ventajas que la media no tiene. Por ejemplo, si queremos saber si en nuestra empresa estamos entre los que cobran más o entre los que cobran menos, debemos comparar nuestro salario con la mediana, y no con la media. Si sólo hay 10 trabajadores y los salarios son (pongamos que en miles de euros): 0,8; 0,8; 0,9; 0,9; 1,0; 1,0; 1,1; 1,1; 1,2 y 10, todos menos 1 (en este caso el 90%) están por debajo de la media, que es 1,88. Esto no pasa nunca con la mediana, si estamos por encima de la mediana, estamos con el 50% de los que más cobran

  Otro ejemplo. Si un examen se aprueba sacando una nota igual o superior a 5 y la nota media que han sacado los estudiantes es de 5, no sabemos cuantos han aprobado. Si se han examinado 50 estudiantes, puede ser que 41 hayan suspendido con un 4; 8 estudiantes hayan sacado un 10 y uno haya obtenido un 6. Esto da media 5, aunque es verdad que son unas notas muy raras. Si la mediana es 5, seguro que la mitad han aprobado.

Además, puestos a criticar la media, podemos decir que su uso conduce a algunas situaciones paradójicas, como aquella que dice que la mayoría de los hombres tiene un número de piernas superior a la media.

Si la distribución de los datos es simétrica, la media y la mediana coinciden, y entonces todo son ventajas. Por ejemplo, en una distribución Normal la media y la mediana son iguales [$P(X>\mu) = 0,5$] y por tanto, si los valores que tenemos provienen de una Normal, y no hay anomalías, la media y la mediana no andarán muy lejos la una de la otra. En cualquier caso, si disponemos de un medio de cálculo fácil, podemos calcular las dos y aprovechar lo mejor de cada una.

2 ¿Tiene alguna aplicación práctica la media geométrica?

Sí la tiene, pero su aplicación es menos frecuente que la media aritmética. Un caso muy típico en que la media geométrica resulta útil es para calcular promedios de tasas de crecimiento. Por ejemplo: una población que tenía 10.000 habitantes en el año cero, creció el primer año a una tasa del 5%, el segundo creció a una tasa del 20% y el tercer año al 50% ¿A qué tasa promedio ha crecido la población en estos 3 años?

Año	Población inicial	Tasa crecimiento	Factor de expansión	Población al final del año
1	10.000	0,05	1,05	10.500
2	10.500	0,20	1,20	12.600
3	12.600	0,50	1,50	18.900

Si calculamos la media aritmética de la tasa de crecimiento tenemos: (0,05 + 0,20 + 0,50) / 3 = 0,25 y el factor medio de expansión sería 1,25. Pero si la población hubiera crecido de esta forma los 3 años, no se llegaría al mismo resultado final:

Año	Población inicial	Tasa crecimiento	Factor de expansión	Población al final del año
1	10.000	0,25	1,25	12.500
2	12.500	0,25	1,25	15.625
3	15.625	0,25	1,25	19.531

Por tanto, la media aritmética no es un buen indicador de la tasa media de crecimiento.

Si la población crece a una tasa constante *i*, para que al final del tercer año tenga el mismo efecto que las tasas del ejemplo, se debe verificar que:

$$10.000(1+i)(1+i)(1+i) = 10.000(1+0,05)(1+0,20)(1+0,50)$$

De donde:

$$(1+i) = \sqrt[3]{1,05 \cdot 1,20 \cdot 1,50} = 1,2364$$

Si se hubiera tenido este factor de expansión cada año (nótese que es la media geométrica), hubiera conducido a una población final exactamente igual a la que tenemos. Es decir, que la tasa media de crecimiento ha sido del 23,64%.

Curiosidades sobre la media geométrica son:

- A diferencia de la media aritmética, la media geométrica solo se define para números positivos.

- La media geométrica nunca es mayor que la media aritmética. La demostración para el caso de 2 valores es fácil por reducción al absurdo. Supongamos que $\sqrt{ab} > (a+b)/2$, entonces $ab > (a^2 + 2ab + b^2)/4$ de donde $0 > a^2 - 2ab + b^2$. Como $a^2 - 2ab + b^2 = (a-b)^2$ es imposible que este valor sea negativo, luego es imposible que $\sqrt{ab} > (a+b)/2$.

Y ya puestos a hablar de otras medias, podemos hacer un comentario sobre la media armónica, mucho menos conocida pero también útil en algunos casos.

Se define la media armónica de $x_1, x_2, \ldots x_N$ como:

$$Mh = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_N}}$$

Parece que esto sea un retorcimiento sin ningún interés, pero no. Por ejemplo, si un coche recorre una cierta distancia a una velocidad media de 100 km/h y vuelve por el mismo camino a 120 km/h, la velocidad media a que ha realizado el viaje es:

$$\frac{2}{\frac{1}{100} + \frac{1}{120}} = 109{,}1\,\text{km/h}$$

y no 110 km/h, como en principio se podría pensar.

Observe que la velocidad es distancia recorrida partido por el tiempo tardado en recorrerla, es decir $v = d/t$ y por tanto $t = d/v$. En nuestro caso, si la distancia a recorrer es *d*, el tiempo tardado en la ida es $t_1 = d/100$ y el tiempo tardado en el regreso es $t_2 = d/120$. De esta manera el tiempo total invertido en todo el recorrido (2*d*) será $t = t_1 + t_2$ y la velocidad media se calcula de la forma:

$$\text{Velocidad media} = \frac{\text{Distancia total recorrida}}{\text{Tiempo total gastado}} = \frac{2d}{\frac{d}{100} + \frac{d}{120}}$$

Cancelando *d* en la expresión anterior se obtiene la fórmula de la media armónica.

Otro ejemplo: Un avión recorre 3.000 km. Los 1000 primeros a 700 km/h, los 1.000 siguientes a 800 km/h, y los 1.000 restantes a 900 km/h ¿Cuál ha sido su velocidad media? No ha sido 800 km/h, sino 791,6 km/h.

3 ¿Por qué en la expresión de la varianza se utiliza el cuadrado de las diferencias en vez de su valor absoluto?

El problema de la varianza es que sus unidades son el cuadrado de las unidades de los datos, y esto dificulta su interpretación. Por eso hacemos su raíz cuadrada, la desviación estándar[1], que es la medida que más usamos para referirnos a la variabilidad.

Claro que podríamos evitar el tener dos medidas –varianza y desviación estándar– utilizando solo una calculada con el módulo (valor absoluto) de la distancia de cada valor con respecto a la media (en vez del cuadrado) y así ya tendría las mismas unidades que los datos. Pero no lo hacemos porque la varianza tiene unas propiedades extraordinarias que ni de lejos presenta esa nueva medida. Vamos a desarrollar unas ideas que nos permitirán justificarlo.

Utilizaremos los datos representados en la Figura 3.1, en la que también hemos representado un valor *a*, en principio arbitrario, con el propósito de descubrir dónde conviene colocarlo para que sea un "buen representante" del conjunto de estos datos. Empezaremos diciendo que *a* puede ser cualquier número real y después le vamos a exigir algunos requisitos asociados con nuestra idea de lo que significa buen representante lo cual restringirá el conjunto de valores que pueda asumir. Veamos 2 criterios para seleccionar el valor de *a*.

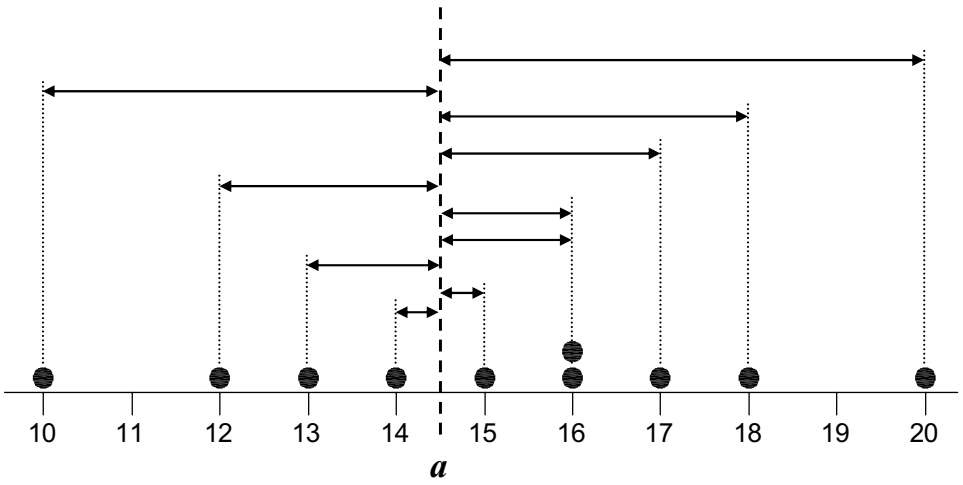


Figura 3.1. *Muestra aleatoria de 10 valores, con sus distancias a un presunto valor central*

Criterio 1

De todos los posibles valores posibles de *a*, escogemos aquel que minimice la función *f*(*a*), definida de la forma:

$$f(a) = \frac{\sum_{i=1}^{N} |x_i - a|}{N}$$

[1] También desviación tipo, o desviación típica.

Los valores x_i, corresponden a los 10 datos que ya conocemos, así que el valor de la distancia promedio depende solo de a, por eso la hemos llamado $f(a)$. El valor que minimiza esta función y que por tanto es el mejor representante de los datos con el criterio aplicado es precisamente su mediana. La demostración de esto no es trivial[2].

Al valor mínimo de $f(a)$ le llamamos desviación media DM con respecto a la mediana, y su fórmula es:

$$DM = \frac{\sum_{i=1}^{N} |x_i - Me|}{N}$$

En nuestro ejemplo, la mediana es 15,5. Esto significa que de todos los números reales, 15,5 es el que está más cerca de los datos de acuerdo con este criterio. Por tanto, la desviación media para nuestros datos es:

$$DM = \frac{|10-15{,}5| + |12-15{,}5| + |13-15{,}5| + |14-15{,}5| + ... + |20-15{,}5|}{10} = 2{,}3$$

Pero el criterio anterior no es el único. Veamos lo que ocurre si aplicamos otro criterio de cercanía.

Criterio 2

De todos los posibles valores de a, vamos a escoger aquel que haga menor la media de los cuadrados de la distancia de los datos a dicho valor a. Es decir, el que minimiza la función:

$$g(a) = \frac{\sum_{i=1}^{N} (x_i - a)^2}{N}$$

En este caso el mejor valor de a puede deducirse derivando $g(a)$ con respecto de a, igualando a cero y despejando su valor. Veamos:

$$\frac{\partial g(a)}{\partial a} = \frac{-2}{N} \sum_{i=1}^{N} (x_i - a) = 0$$

Por tanto $\sum_{i=1}^{N} (x_i - a) = 0$, de donde se deduce que $\sum x_i = N \cdot a$ y despejando a tenemos: $a = \sum_{i=1}^{n} x_i / N = \mu$

[2] La demostración puede encontrarse en el texto de Enrique Cansado: *Estadística General*. Centro Internacional de Enseñanza de la Estadística (CIENES). Santiago de Chile, 1967.

Si hacemos la segunda derivada vemos que siempre es positiva, lo cual confirma que el punto crítico es $a = \mu$ (media aritmética) y que el valor mínimo de $g(a)$ es la varianza de X.

Con los datos de nuestro ejemplo $\mu = 15{,}1$ y el valor mínimo que toma $g(a)$ es $\sigma^2 = 7{,}89$. Sacando raíz cuadrada se obtiene la llamada desviación estándar $\sigma = 2{,}81$.

¿Por qué se prefiere usar la varianza, deducida a través del criterio 2, en lugar de la desviación media, deducida aplicando el criterio 1? Veamos algunas razones:

1. Los desarrollos anteriores ponen de manifiesto que la media, medida descriptiva por excelencia, está asociada de forma más natural con la varianza que con la *DM*.

2. La *DM*, ya sea respecto a la mediana o respecto a la media, incluye en su expresión la función "valor absoluto" que se comporta mal desde un punto de vista matemático, mientras que la función cuadrática que implica la suma de cuadrados de las desviaciones, es muy fácilmente tratable. Observe, por ejemplo, que la demostración de que la media hace mínimo el promedio de los cuadrados se ha realizado de forma casi inmediata, mientras que probar que la mediana hace mínima la media de las distancias es bastante más complejo.

3. De forma natural, la desviación estándar σ de la población, forma parte de la distribución más famosa y útil, como es la Normal.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

   Esto posibilita la construcción de intervalos de confianza para estimar, por ejemplo, la media de la población, lo que sería mucho más complejo si se usara la *DM*, en lugar de σ.

4. La varianza es una suma de cuadrados que se puede descomponer en diversos sumandos, dando origen al llamado Análisis de la Varianza. El desarrollo de la teoría de los modelos lineales está basado en gran parte en el criterio de los mínimos cuadrados, que es el mismo con que se ha obtenido la media. Cuando la población origen de los datos es Normal, el cociente de varianzas de muestras independientes sigue una distribución conocida, llamada F de Snedecor.

5. La varianza de una suma de variables aleatorias se calcula de una forma muy fácil, especialmente si las variables son independientes. Por ejemplo, en un proceso de envasado, si el peso del envase tiene una varianza $V(X)$ y la varianza del contenido es $V(Y)$, la varianza del conjunto es $V(X+Y) = V(X) + V(Y)$.

Nada de esto podríamos hacer si intentamos usar la desviación media como medida de dispersión. En síntesis, la teoría estadística desarrollada con base en la varianza es muy rica, y no se conoce nada parecido para la desviación media.

4 ¿Por qué cuando se calcula la varianza de una muestra se divide por *n*-1 en vez de dividir por *n*?

Para hacer más comprensible la explicación vamos a trabajar con un ejemplo suponiendo que conocemos la población completa (lujo que no tendremos en la práctica). Las unidades que componen la población son: A, B, C, D, E, F y sus mediciones respectivas son:

(A)	(B)	(C)	(D)	(E)	(F)
2	6	8	10	10	12

En primer lugar ilustraremos la propiedad de insesgamiento de un estimador para el caso de la media, con la cual estamos más familiarizados, y luego repetiremos la experiencia para el caso que nos ocupa de la varianza.

La media poblacional μ, de los datos del ejemplo, es:

$$\mu = \frac{2+6+8+10+10+12}{6} = 8$$

Supongamos que queremos estimar ("hacernos una idea") el valor μ**,** usando una muestra aleatoria de $n = 2$ unidades de la población. En este caso, en que la población consta solo de $N = 6$ unidades, podemos hacer un listado de todas las posibles muestras que pudieran resultar al escoger dos unidades al azar. Estas muestras aparecen enumeradas en la Tabla 4.1.

Tabla 4.1. *Listado de todas las posibles muestras con n=2 que resultarían en una elección al azar*

Muestra Nº	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Unidades de la muestra	A B	A C	A D	A E	A F	B C	B D	B E	B F	C D	C E	C F	D E	D F	E F
Valores en la muestra	2 6	2 8	2 10	2 10	2 12	6 8	6 10	6 10	6 12	8 10	8 10	8 12	10 10	10 12	10 12
Media muestral $\bar{x}$	4	5	6	6	7	7	8	8	9	9	9	10	10	11	11

Cuando seleccione al azar una muestra de dos unidades, el resultado será necesariamente alguna de estas 15 posibles combinaciones de 2 elementos, con su media $\bar{x}$ correspondiente.

Decimos que $\bar{x}$ es un estimador insesgado de μ, si el promedio de todas las posibles medias coincide exactamente con la media de la población. Para verificarlo, hagamos el promedio de nuestras 15 posibles medias:

$$\frac{4+5+6+6+7+7+8+8+9+9+9+10+10+11+11}{15} = 8 \equiv \mu$$

El promedio coincide perfectamente con μ. Esto pasa en todos los casos, independientemente del tipo de población o del tamaño de la muestra, por eso decimos que $\overline{x}$ es un estimador insesgado para μ.

Veamos ahora si el estadístico

$$S_n^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$$

donde *n* es el tamaño de las muestras, es un estimador insesgado para la varianza σ_N^2 calculada como

$$\sigma_N^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

donde N es el tamaño de la población. Queremos saber si el promedio de los valores de S_n^2, para cada una de las posibles muestras, da el valor de σ_N^2 y para averiguarlo, en primer lugar vamos a calcular la varianza poblacional:

$$\sigma_N^2 = \frac{(2-8)^2 + (6-8)^2 + (8-8)^2 + (10-8)^2 + (10-8)^2 + (12-8)^2}{6} = \frac{64}{6} = 10,67$$

Ahora calculamos la varianza para cada muestra de 2 unidades, con la fórmula:

$$S_n^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2}{2}$$

obteniéndose los resultados que aparecen en la Tabla 4.2.

Tabla 4.2. *Varianza muestral para cada una de las 15 posibles muestras de tamaño n=2*

Muestra Nº	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Valores en la muestra	2 6	2 8	2 10	2 10	2 12	6 8	6 10	6 10	6 12	8 10	8 10	8 12	10 10	10 12	10 12
Varianza muestral S^2	4	9	16	16	25	1	4	4	9	1	1	4	0	1	1

Veamos si el promedio de las posibles varianzas muestrales, coincide con 10,67 que es el valor obtenido para σ_N^2.

$$\overline{S}^2 = \frac{4+9+16+16+25+1+4+4+9+1+1+4+0+1+1}{15} = \frac{96}{15} = 6,4 \neq {\sigma_N}^2$$

No coincide y, por tanto, S_n^2 no es un estimador insesgado de σ_N^2.

Sin embargo S_{n-1}^2, definido de la forma:

$$S_{n-1}^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

aunque tampoco es un estimador insesgado para σ_N^2, sí lo es para σ_{N-1}^2 definido como:

$$\sigma_{N-1}^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N-1}$$

Efectivamente, σ_{N-1}^2 tiene el valor:

$$\sigma_{N-1}^2 = \frac{(2-8)^2+(6-8)^2+(8-8)^2+(10-8)^2+(10-8)^2+(12-8)^2}{6-1} = \frac{64}{5} = 12,8$$

Y si calculamos S_{n-1}^2 para cada una de nuestras muestras deberemos aplicar la fórmula:

$$S_{n-1}^2 = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2}{2-1} = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2}{1}$$

Es decir, que todos los valores de la varianza que aparecen en la Tabla 4.2, quedan ahora multiplicados por 2 y por lo tanto la media de las varianzas queda también multiplicada por 2, es decir:

$$\overline{S}_{n-1}^2 = 2 \cdot \frac{96}{15} = 2 \cdot 6,4 = 12,8 = \sigma_{N-1}^2$$

Quizá este resultado decepcione, y hasta sorprenda, porque seguramente lo esperado era que S_{n-1}^2 fuera un estimador insesgado de σ_N^2 y no de σ_{N-1}^2, pero no hay que preocuparse demasiado. A efectos prácticos es casi lo mismo cuando la población es grande, y nosotros no vamos a estimar a través de muestras las características de una población de 6 elementos (en nuestro ejemplo esta ha sido una población "de juguete" para entender lo que estábamos haciendo). Cuando estimemos la varianza de una población se tratará de una población grande, en la que σ_N^2 será prácticamente igual a

σ^2_{N-1}. En realidad, el caso más frecuente es tener poblaciones teóricas (infinitas), en las que es exactamente lo mismo σ^2_N que σ^2_{N-1}.

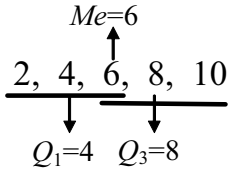
Una observación final

Es importante notar que incluso cuando la población es infinita S_{n-1} no es un estimador insesgado de ⇑. La raíz cuadrada estropea las propiedades que sí tiene el estimador de la varianza. Se incluye otro comentario sobre este particular al final de la respuesta a la pregunta 49.

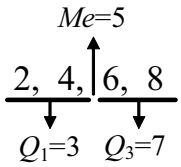
5 ¿Cuál es la forma "correcta" de calcular los cuartiles?

Suele generar confusión, y también curiosidad, comprobar que se utilizan diferentes criterios para determinar los cuartiles. Veamos algunos ejemplos.

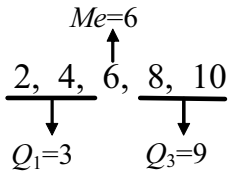
John Tukey, creador de muchas de las modernas técnicas de análisis exploratorio de datos, en su libro *Exploratory Data Analysis* construye los *boxplots* a partir de unos cuartiles que identifica buscando las medianas de los valores que quedan por encima y por debajo de la mediana global. En la determinación de las medianas, de cada mitad de datos incluye la mediana global si el número total de datos es impar, pero no la incluye si es par. Por ejemplo, para los datos 2, 4, 6, 8 y 10 tenemos:



y si los datos son 2, 4, 6, y 8 (número par):



David Moore y George McCabe en su libro *Introduction to the Practice of Statistics*, muy conocido y valorado por su carácter pedagógico e innovador, proponen un método similar al de Tukey pero sin incluir la mediana global a la hora de determinar las medianas de cada una de las mitades. Cuando el número de datos es par el valor de los cuartiles coincide con el método de Tukey, pero en general no coincide cuando el número de datos es impar:



El *software* estadístico Minitab utiliza las expresiones $0{,}25(n+1)$ y $0{,}75(n+1)$ para determinar las posiciones de Q_1 y Q_3 respectivamente. Si la posición obtenida para Q_1 es, por ejemplo, 1,25, su valor estará comprendido entre el primero (x_1) y el segundo (x_2) interpolando la forma: $Q_1 = x_1 + 0{,}25(x_2 - x_1)$. Utilizando los mismos datos que en los ejemplos anteriores resulta.

Datos: 2, 4, 6, 8

Posición de Q_1: 0,25 × 5 = 1,25
Valor de Q_1 = 2 + 0,25 (4-2) = 2,5

Posición de Q_3: 0,75 × 5 = 3,75
Valor de Q_3: 6 + 0,75 (8-6) = 7,5

Datos 2, 4, 6, 8, 10

Posición de Q_1: 0,25 × 6 = 1,5
Valor de Q_1 = 2 + 0,5 (4-2) = 3

Posición de Q_3: 0,75 × 6 = 4,5
Valor de Q_3: 8 + 0,5 (10-8) = 9

Excel identifica las posiciones de los cuartiles mediante las expresiones: 0,25(n-1)+1 y 0,75(n-1)+1 interpolando de la misma forma que hemos visto para Minitab. Con nuestros datos se obtienen los valores que se indican en la Figura 5.1.

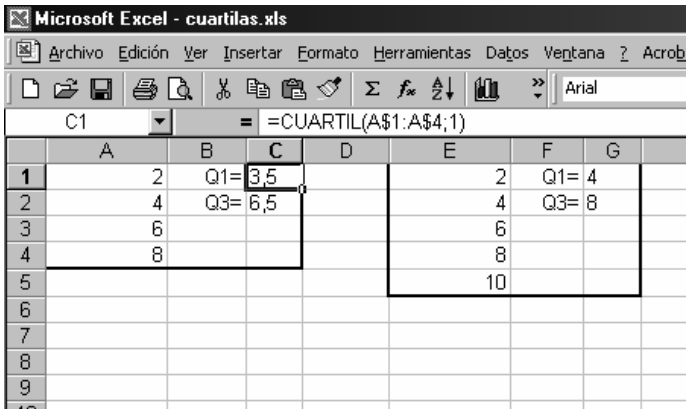


Figura 5.1. *Determinación de los cuartiles con Microsoft Excel*

En resumen, los valores obtenidos para los cuartiles con los métodos comentados han sido:

Método	Datos: 2, 4, 6, 8		Datos: 2, 4, 6, 8, 10	
	Q_1	Q_3	Q_1	Q_3
Tuckey	3	7	4	8
Moore y McCabe	3	7	3	9
Minitab	2,5	7,5	3	9
Excel	3,5	6,5	4	8

¿Cuál es el método correcto? ¿Cuál debemos utilizar? La verdad es que en la práctica no importa demasiado. Cuando se está interesado en conocer el valor de los cuartiles el conjunto de datos es grande, y en este caso las diferencias entre los distintos métodos de cálculo son pequeñas.

Por ejemplo, si tenemos 500 valores, la posición para el primer cuartil es 125,25 si se usa el criterio de Minitab, y 125,75 si se usa el de Excel. Lo normal es que haya muy poca diferencia entre el valor que ocupa la posición 125 y el que ocupa la 126 y, por tanto, la diferencia según se aplique un criterio u otro será poco importante a efectos prácticos.

6 ¿En cuántos intervalos conviene dividir los datos para construir un histograma? ¿Qué otros aspectos hay que tener en cuenta?

Respecto al número de intervalos no hay una regla fija, aunque lo razonable es que su número aumente al ir aumentando el número de datos. Si se utiliza un programa de ordenador, este ya dará un número de intervalos razonable. Si se hace a mano, una regla sencilla para tomar como referencia es la siguiente:

Núm. de datos	Núm. de intervalos
20* – 50	7
50 – 75	10
75 – 100	12
Más de 100	15

*Para menos de 20 datos es mejor utilizar un diagrama de puntos

Pero también debe tenerse en cuenta que para facilitar la lectura del histograma es importante que la anchura de los intervalos sea un número sencillo. Por tanto, la tabla anterior se debe utilizar como primera aproximación, ya que el número exacto estará supeditado a tener un valor adecuado para la anchura de los intervalos.

Veamos a través de un ejemplo los aspectos más relevantes a tener en cuenta, tanto si los histogramas se contruyen con ordenador como si se hacen a mano. La Tabla 6.1 contiene los pesos (en gramos) de 160 porciones de mantequilla, 80 cortados y empaquetados con la máquina 1 y otras 80 con la máquina 2. El valor nominal es de 220 gramos, se considera tolerable una desviación de ± 10 gramos y existe interés en conocer y comparar la variabilidad que presentan los pesos en ambas máquinas.

Tabla 6.1. *Datos correspondientes al peso, en gramos, de 160 porciones de mantequilla*

Máquina 1				Máquina 2			
220,3	215,5	219,1	219,2	220,3	208,0	214,4	219,2
215,8	222,0	218,9	213,6	216,9	213,4	217,7	217,7
220,4	218,7	218,6	219,6	222,9	219,7	209,4	221,6
221,5	227,0	219,5	222,5	223,1	215,3	220,4	215,6
215,7	225,3	223,0	218,0	216,0	210,9	221,4	210,9
222,7	215,1	219,6	217,3	212,1	213,0	218,0	216,5
216,0	218,8	217,9	213,0	216,9	216,0	213,5	219,2
219,4	218,3	216,7	224,1	216,2	218,4	216,6	214,9
219,8	222,6	219,1	217,7	216,2	212,2	216,9	214,9
220,2	219,5	222,4	219,9	222,9	214,3	219,1	216,7
218,0	223,9	219,6	221,9	214,9	212,6	219,4	213,3
219,3	219,6	218,8	219,9	219,0	216,7	216,4	213,5
220,0	214,1	224,3	217,4	218,0	219,5	219,5	222,3
223,9	220,6	219,5	219,6	211,8	218,2	218,3	217,4
218,1	218,8	218,4	217,9	214,6	215,7	218,0	216,4
216,9	221,6	220,6	222,6	215,6	220,4	217,3	216,2
217,9	225,7	222,2	216,1	212,5	214,6	209,7	211,3
224,2	216,2	219,9	220,4	215,8	219,9	216,5	211,9
214,1	219,7	222,4	224,5	213,7	209,7	216,9	213,1
221,1	225,0	222,7	222,2	212,5	217,5	217,4	215,7

La Figura 6.1 muestra los histogramas construidos con Excel (Excel 2000: *Herramientas > Análisis de datos > Histograma*) con los valores de las escalas y de anchura de las barras que se tienen por defecto. En ambos casos aparecen 9 barras (que deberían tocarse, ya que la variable representada es continua), pero lo más destacable es que los números que figuran en el eje horizontal son "raros", especialmente para la máquina 2, y esto dificulta su lectura y la interpretación del gráfico. Además, como todos los programas adaptan la escala al rango de variación de los datos, las escalas no son iguales para las 2 máquinas, lo que complica la comparación.

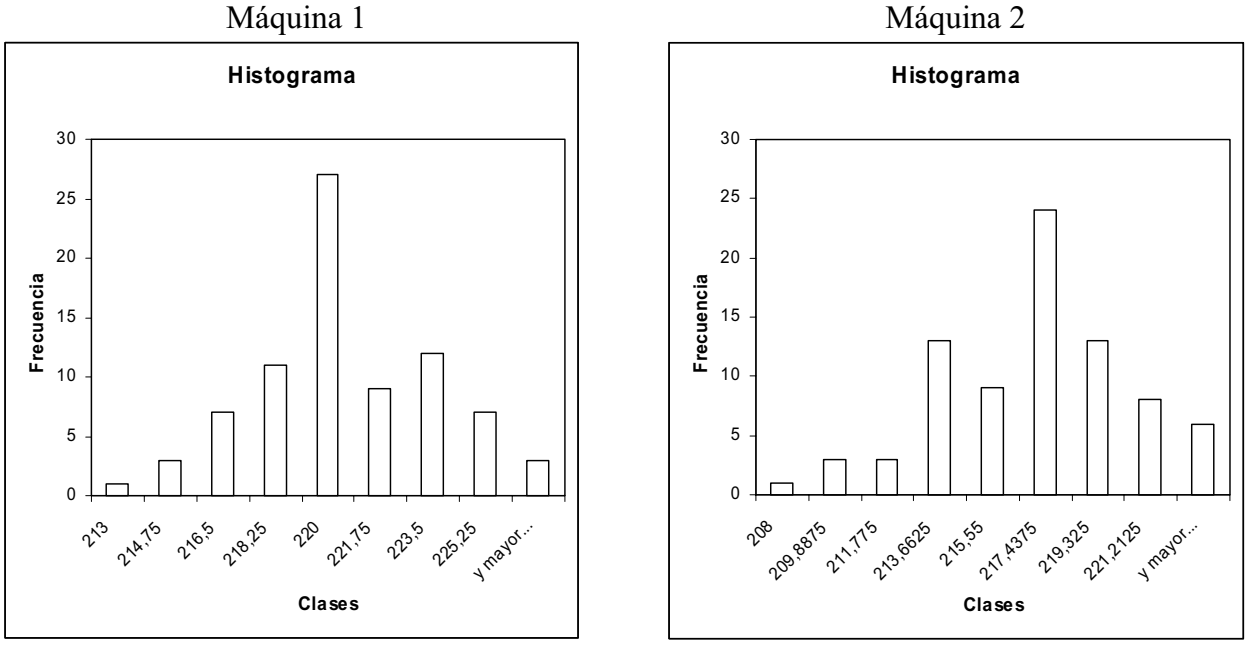


Figura 6.1. *Histograma construido con Excel con todos los parámetros por defecto*

Si utilizamos Minitab (Versión 13: *Graph > Histogram*) con todos los parámetros por defecto, aparecen 15 barras para la máquina 1 y 16 para la 2. En este caso, tanto los números que figuran en los ejes como la anchura de los intervalos (1 gramo), son fáciles de leer, aunque sería mejor tener más valores en el eje horizontal. En cuanto a las escalas, ocurre lo mismo que en el caso anterior.

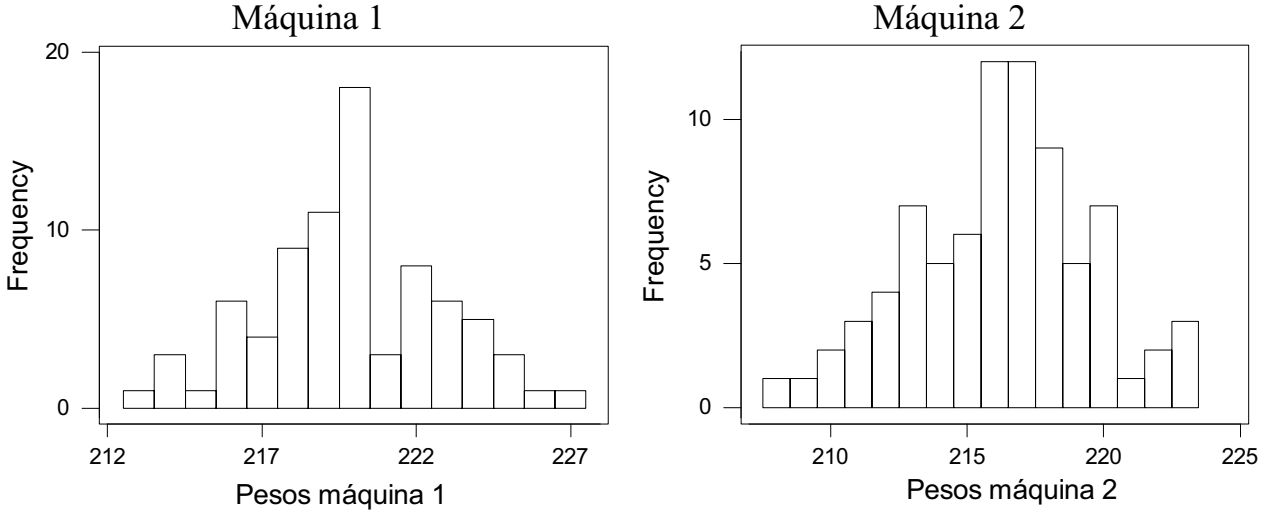


Figura 6.2. *Histogramas construido con Minitab con los parámetros por defecto*

Actuando sobre las opciones de Minitab se han construido los histogramas de la Figura 6.3, en los que los ejes están marcados con números que facilitan la lectura, se ha mantenido una anchura de intervalo de 1 gramo y se ha forzado que las escalas sean iguales. También se han añadido unas líneas con el valor nominal y las tolerancias, de forma que con solo dar un vistazo se observa que la máquina 1 está produciendo básicamente bien, mientras que la 2 está descentrada.

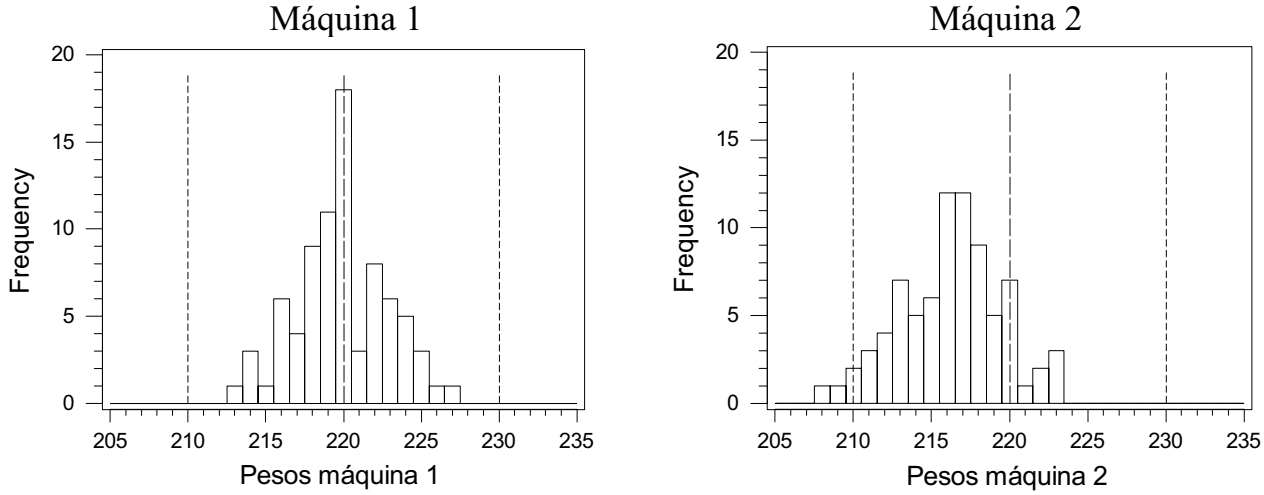


Figura 6.3. *Histogramas construidos con Minitab actuando sobre las opciones disponibles para conseguir la apariencia deseada*

En resumen, si el histograma se construye con un programa de ordenador, los aspectos a tener en cuenta para facilitar su lectura e interpretación son:

- Los ejes, especialmente el horizontal, deben estar marcados con valores fáciles de leer.
- La anchura de los intervalos conviene que sea también un número "redondo".
- Si se van a comparar varios histogramas, es necesario que todos ellos estén construidos con la misma escala para facilitar la comparación y evitar confusiones.

Si estas características no aparecen con los parámetros que el programa tiene configurados por defecto, conviene actuar sobre las opciones disponibles para conseguirlo.

¿Y si se hace a mano? En este caso los pasos a seguir son:

1. Calcular el rango de los datos (valor máximo menos valor mínimo). En el caso de la máquina 1, $R = 227{,}0 - 213{,}0 = 14{,}0$.
2. Plantear un número de intervalos en primera aproximación. En nuestro caso, con 80 datos, la tabla guía indica $k = 12$ intervalos.
3. Calcular la anchura del intervalo, h, y ajustar a un número redondo. $h = R/k$, en nuestro caso $h = 14/12 = 1{,}17$, y por tanto lo más razonable es redondear a 1.

4. Tabular los datos de acuerdo con los intervalos definidos. Tener en cuenta que también interesa que los límites de los intervalos, o la marca de clase, sean números sencillos.

5. Construir el histograma. Si se va a comparar con otros, la escala debe ser lo suficiente amplia para que pueda ser común a todos ellos, y también conviene mantener en común tanto la anchura de los intervalos como sus extremos.

Una consideración para terminar. Seguramente estaremos de acuerdo en que construir histogramas a mano es una tarea un tanto tediosa, y si hacerlos con ordenador no es posible o implica gestiones y retrasos que no compensan, una buena idea puede ser utilizar una plantilla para recoger los datos de forma que al irlos anotando el histograma se vaya construyendo solo, tal como se indica, a título de ejemplo, en la Figura 6.4. Incluso aunque se tenga ordenador disponible, la inmediatez en el análisis de los datos que se obtiene con este método puede hacer que sea el más adecuado.

Naturalmente, hay que tener una idea de por dónde irá la variabilidad de los datos para poder diseñar la plantilla, que además, para su completa identificación y posibles análisis comparativos, siempre debe incluir un apartado con la fecha, el origen de los datos, la persona que los tomó, etc.

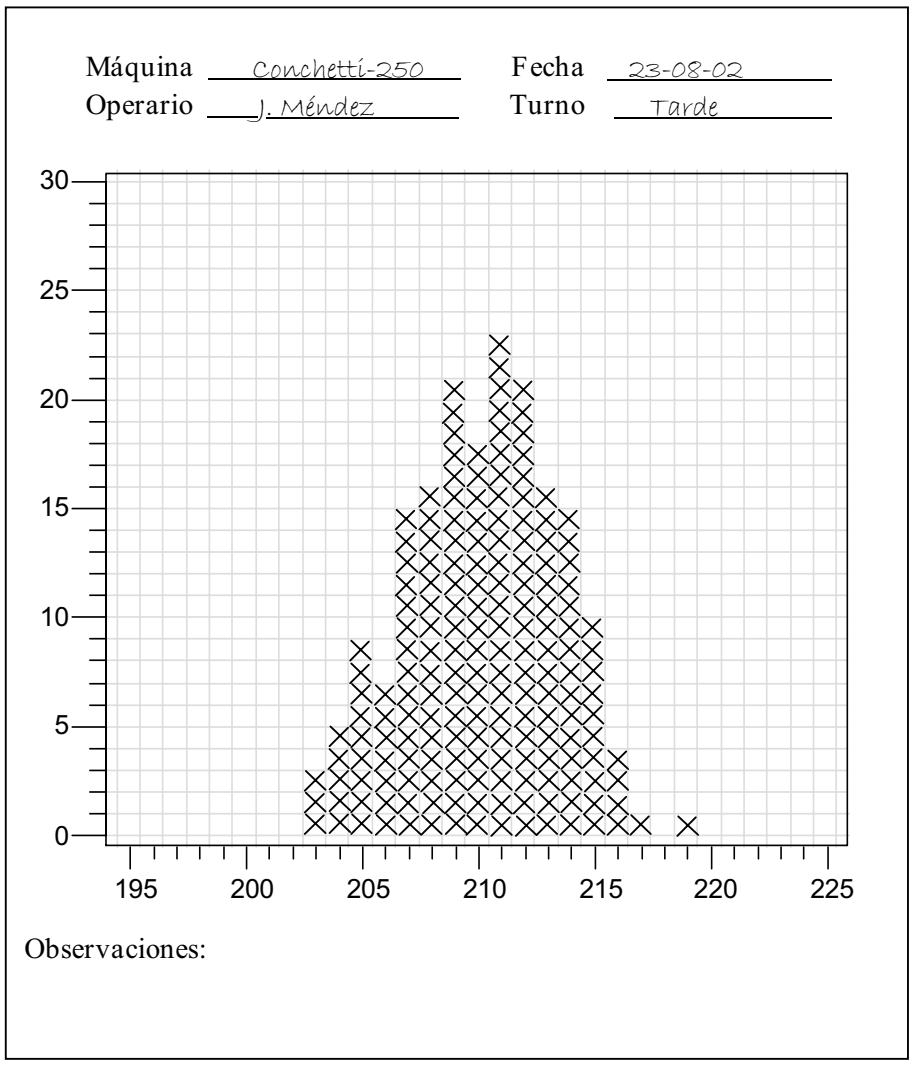


Figura 6.4. *Plantilla de recogida de datos en la que el histograma se va construyendo solo. Los valores se redondean a las unidades y se marca una cruz en el lugar correspondiente*

7 ¿Cuándo conviene utilizar *boxplots* para analizar o describir datos?

Los *boxplots* son gráficos muy apropiados para mostrar el comportamiento de los datos cuando interesa presentarlos estratificados por alguna variable cualitativa. Por ejemplo, la Figura 7.1 muestra la distribución de los pesos de 500 paquetes de azúcar llenados en una planta de envasado que consta de 5 líneas independientes. Los pesos se presentan según la línea en que se han llenado (100 paquetes por línea). El valor nominal es de 1.000 g.

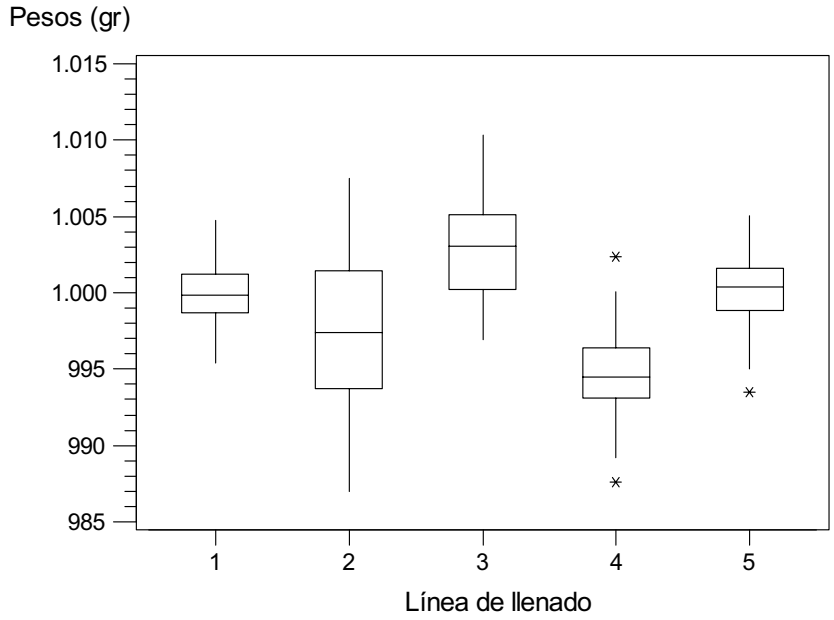


Figura 7.1. *Distribución de los pesos de paquetes de azúcar según la línea en que han sido llenados*

Fácilmente se puede observar que en las líneas 1 y 5 la producción está centrada en el valor objetivo con una variabilidad de aproximadamente ± 5 *g*, mientras que las otras líneas están descentradas, presentando la línea 2 una variabilidad claramente mayor que las otras.

Otra situación en la que el uso de *boxplots* resulta adecuado es la mostrada en la Figura 7.2, que representa los precios de venta de los pisos en Barcelona en 1990 (en miles de pesetas por metro cuadrado) según el distrito en que se encontraban[1].

En ambos casos los gráficos resumen de una forma clara y compacta la información que contienen los datos. Además facilitan la comparación entre grupos y permiten una rápida identificación de valores atípicos.

[1] Los datos provienen de un estudio realizado por el "Centre de Política del Sòl i Valoracions" del departamento de Construcciones Arquitectónicas I de la UPC.

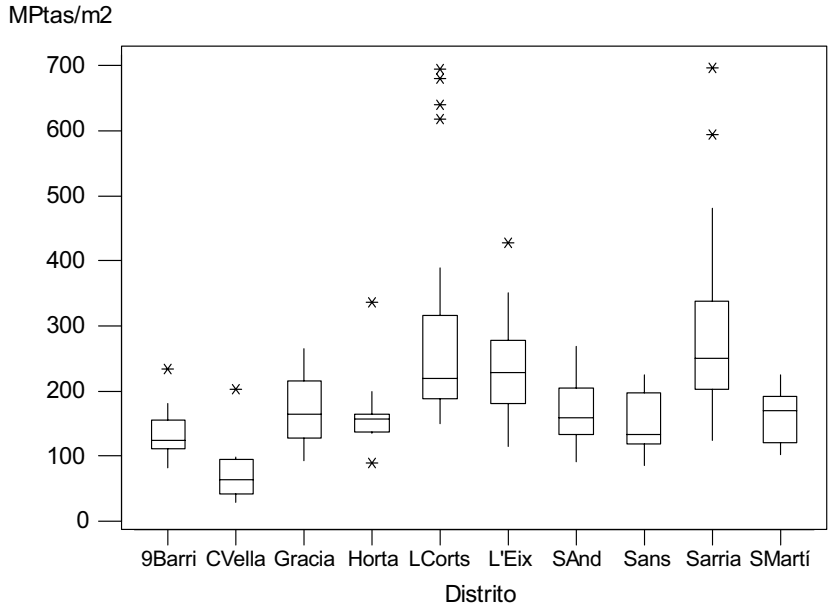
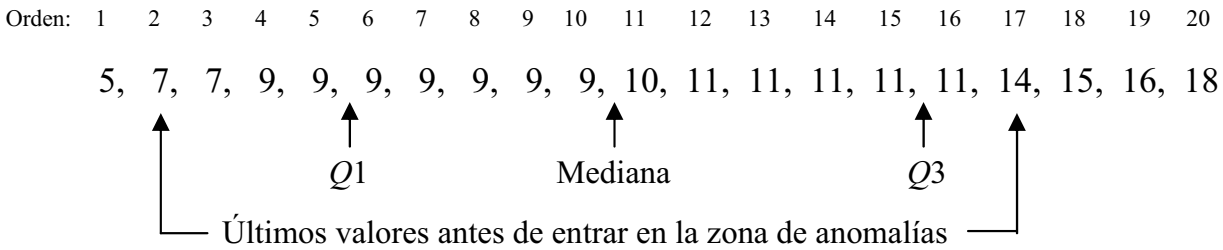
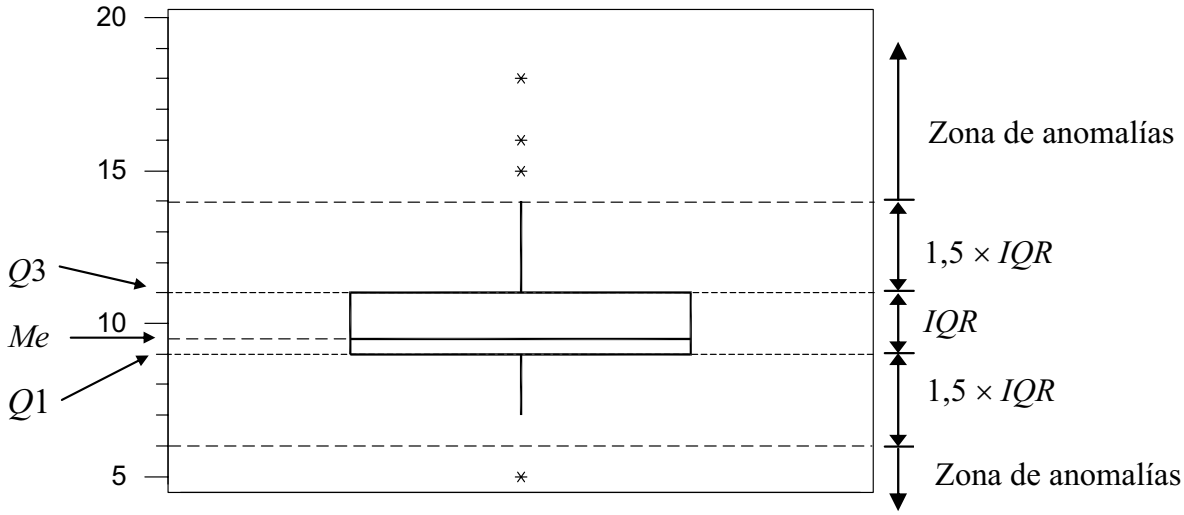


Figura 7.2. *Precio de venta de los pisos en Barcelona en 1990 (en miles de pesetas por metro cuadrado) según el distrito en que se encontraban*

Anexo: Ejemplo de construcción de *boxplot*





Siendo: *Q1*: Primer cuartil; *Q3*: Tercer cuartil; *IQR*: Rango intercuartílico (*Q*3-*Q*1).

8 En los *boxplots* las anomalías se marcan a partir de ± 1,5 veces el rango intercuartílico (IQR) ¿De dónde sale el 1,5?

Efectivamente, este parece un número caprichoso. En principio también el 1 o el 2 podrían ser buenos candidatos, con la ventaja de que son números más sencillos, pero veamos qué ocurriría si fueran estos los elegidos.

En primer lugar debemos tener en cuenta que los valores que aparecen en la zona de anomalías de un *boxplot* tienen sentido como tales anomalías (o, quizá mejor dicho "presuntas anomalías") si la población de la que provienen los datos es Normal. Consideremos por tanto una distribución Normal, y puestos a elegir una tomaremos la $N(0;1)$ ya que esto no va a restringir las conclusiones a las que vamos a llegar. Para determinar los cuartiles buscamos el valor de z que deja un área de cola de 0,25 y resulta ser: $z_{0,25} = 0,674$. Por tanto, el rango intercuartílico será: $IQR = 0,674 - (-0,674) = 1,348$.

Si la zona de anomalías empieza en $Q1 - 1 \times IQR$ y $Q3 + 1 \times IQR$, el valor correspondiente al lado derecho será: $Q3 + 1 \times IQR = 0,674 + 1,348 = 2,022$, y $P(Z>2,022) = 0,02$. Por tanto, la probabilidad de que un valor que pertenezca a la distribución considerada aparezca en la zona de anomalías es del 4% (2% por cada lado).

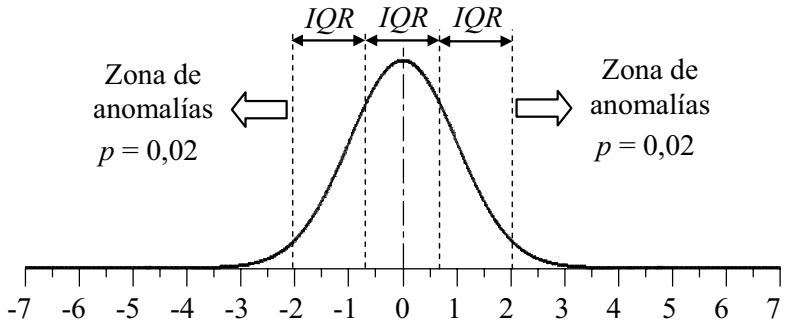


Figura 8.1. *Zona de anomalías con el criterio de cuartiles ± 1·IQR*

Si tomamos el valor 2 como multiplicador del rango intercuartílico para determinar el inicio de la zona de anomalías, tendremos que: $Q3 + 2 \times IQR = 0,674 + 2 \times 1,348 = 3,37$ y $P(Z>3,37) = 0,0004$, por lo que la probabilidad de que un valor aparezca como anomalía es de 0,08%.

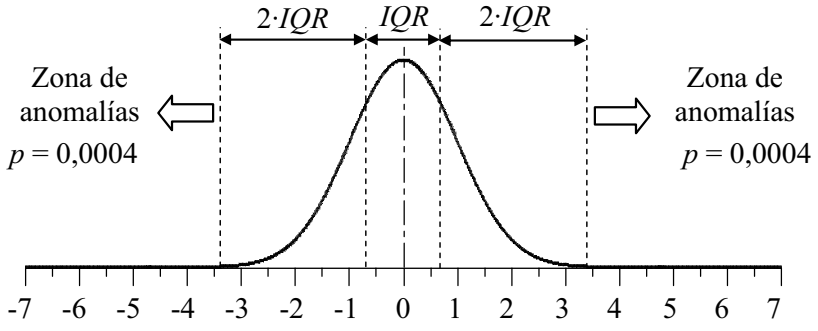


Figura 8.2. *Zona de anomalías con el criterio de cuartiles ± 2·IQR*

John Tukey, que fue el primero en plantear el uso de los *boxplots* para el análisis gráfico de datos, ya contempló la posibilidad de usar los valores 1 o 2, pero consideró que 1 es demasiado pequeño ya que la zona de anomalías con este criterio incluye valores que no merecen ser considerados como tales, y el valor 2 resulta excesivamente grande ya que aleja demasiado esta zona y por tanto pueden pasar desapercibidos valores que deben ser considerados como anómalos.

Descartado el 1 y el 2, y estando claro que el más adecuado es un valor entre ellos, aparece la opción del 1,5 como número más sencillo, y al definir una zona de anomalías con probabilidades muy razonables (p = 0,007) este fue el valor que se propuso y así se ha consolidado.

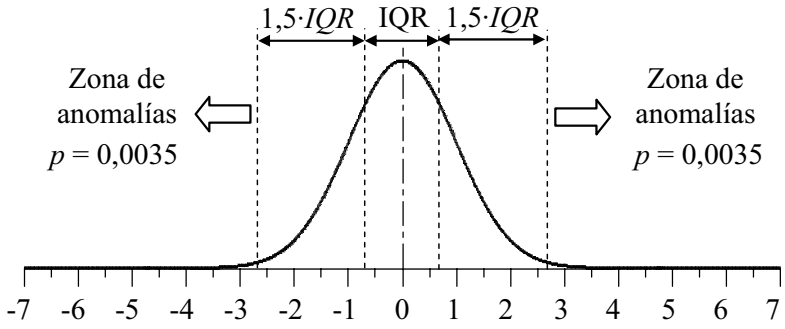


Figura 8.3. *Zona de anomalías con el criterio de ± 1,5·IQR*

9 ¿Qué hay que hacer cuando nos encontramos con valores atípicos?

Para empezar podemos decir que hay 2 cosas que NO DEBEN HACERSE: 1) ignorarlos siempre como si no existieran; o 2) eliminarlos inmediatamente sin más consideraciones.

En la práctica el error más frecuente consiste en no preocuparse por su posible existencia, lanzándose directamente a realizar el test que corresponda. Este modo de proceder tiene riesgos importantes, como el de trabajar con valores que podrían ser erróneos (ya sea porque se han introducido mal en el ordenador, porque están en unidades que no corresponden, ...etc.) y el incluirlos en el estudio puede conducir a unos resultados no válidos.

También puede darse el caso en que los valores sean correctos pero no convenga tomarlos en consideración. Por ejemplo, supongamos que se recogen datos para ver si se confirma la sospecha de que un coche de policía aparcado en la orilla de la carretera recuerda a los conductores cuál es la velocidad máxima en ese tramo. Para ello, durante cierto periodo de tiempo se mide, con un radar oculto, la velocidad de los vehículos en una zona en la que la máxima permitida es de 60 km/h. Las velocidades obtenidas (en km/h) son:

65	66	80	57	57	74	55	59	65	60	77	72	63	55	74	67	63
25	57	20	68	22	61	59	77	72	23	67	58	66	71	71	56	63

A continuación se toman las mismas mediciones pero colocando un coche de la policía aparcado al lado de la carretera de forma visible, en este caso las velocidades son:

55	55	63	58	62	57	59	70	22	58	56	63	55	61	58	60	24	55	75
58	61	63	20	63	62	55	25	80	61	61	60	59	73	60	58	60	58	57

Si realizamos el test de *t* de Student para muestras independientes contrastando la hipótesis nula de que las velocidades medias son iguales frente a la alternativa de que con el coche de la policía son menores, se obtiene un p-valor de 0,167, por lo que del estudio no se podría deducir que el coche de policía tiene efecto disuasorio.

Sin embargo, realizando el análisis exploratorio de los datos, se obtiene un gráfico como el de la Figura 9.1 en el que se observan unos valores atípicos, tanto con coche de la policía como sin él, que corresponden a vehículos que han pasado a unos 20-25 km/h. ¿Qué hay que hacer con estos valores? Lo primero es preguntarse a qué corresponden, cómo es que se han producido. En este caso, después de una simple reflexión, se llega a la conclusión de que corresponden a vehículos de transporte agrícola que siempre van a esta velocidad, porque no pueden ir a más. Está claro que la posible influencia del método disuasorio no va con este tipo de vehículos y lo más razonable es excluirlos del estudio.

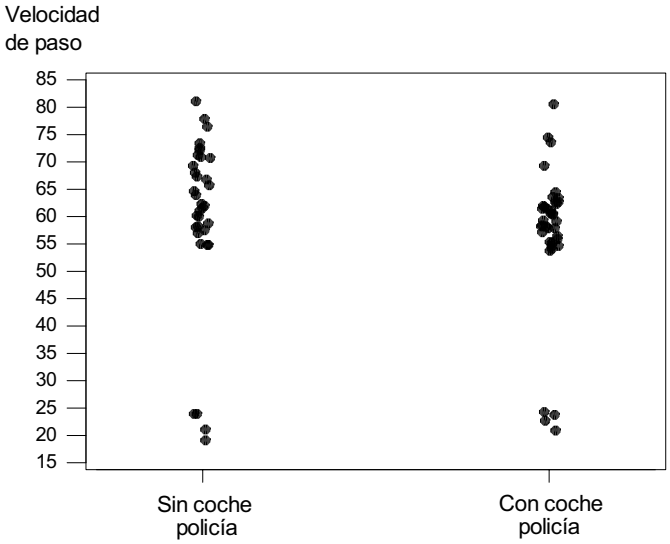


Figura 9.1. *Velocidades de paso de los coches sin y con coche de policía a la vista*[1]

Al eliminar estos valores el p-valor del test de comparación de medias pasa a ser 0,006, por lo que la conclusión sería que el efecto disuasorio del coche de policía es eficaz.

Pero tampoco hay que caer en la tentación de eliminar las anomalías automáticamente siguiendo la política del "aquí te pillo aquí te mato". Si los valores obtenidos hubieran sido los que se reflejan en la Figura 9.2, que son igual a los anteriores añadiendo 115, 118, 120 y 117 a la velocidad de paso sin coche de policía, tendríamos en este grupo 2 conjuntos de valores atípicos, pero aunque hemos visto que lo razonable es quitar el conjunto de los valores bajos, no hay ninguna razón para quitar el de los valores altos, que corresponden a coches que circulan a una velocidad mucho mayor a la permitida y para los que el coche de la policía parece ser un método de disuasión eficaz.

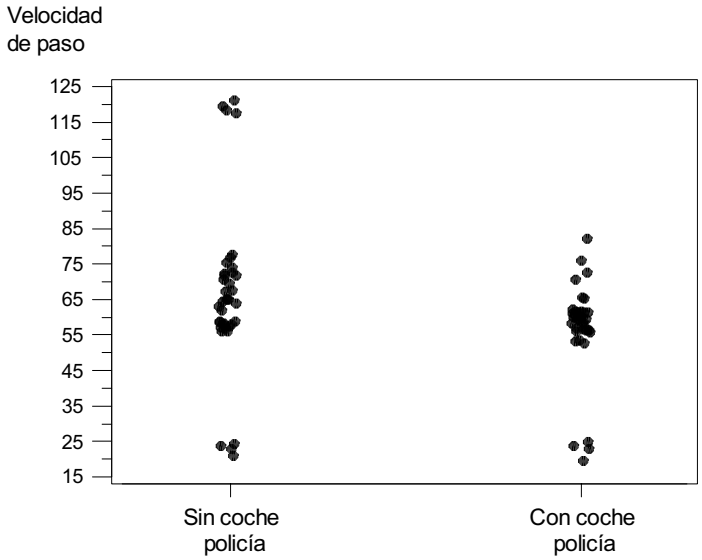


Figura 9.2. *Velocidades de paso con valores atípicos por arriba y por abajo*[2]

[1] Se ha realizado con el *software* estadístico Minitab utilizando la opción *Add Jitter* que varía ligeramente las coordenadas de los puntos para evitar que aparezcan superpuestos y se pierda información sobre su densidad.

[2] Al haberse aplicado la opción *Add Jitter* las coordenadas de los puntos no son idénticas en ambos gráficos.

Otro aspecto a tener en cuenta es que en algunas situaciones, la identificación y el análisis de las anomalías es la parte más interesante del estudio y de la que más frutos se pueden obtener. Por ejemplo, el gráfico de la Figura 9.3 muestra el análisis de unos datos que ponen de manifiesto la relación entre el rendimiento obtenido en una reacción química y la temperatura a que ha sido realizada. En el gráfico aparecen unos valores claramente anómalos, los que están en torno a 185 °C y 77-78% de rendimiento y el que está a 205 °C y 71% de rendimiento. ¿Qué hacer con estos valores? ¿Eliminarlos y olvidarse de ellos?

Si lo hiciéramos así nos perderíamos la oportunidad de incorporar información valiosa a nuestro conocimiento del proceso. Lo más adecuado sería preguntarnos: ¿Por qué se han dado estas situaciones?, ¿qué ha ocurrido a 185 °C para que se hayan producido unos rendimientos tan anormalmente altos?, ¿por qué una vez se hizo la reacción a 205 °C y se obtuvo un rendimiento tan bajo? Es posible que la respuesta a estas preguntas nos aporte información que puede ser muy útil para nuestro mayor dominio y conocimiento del proceso.

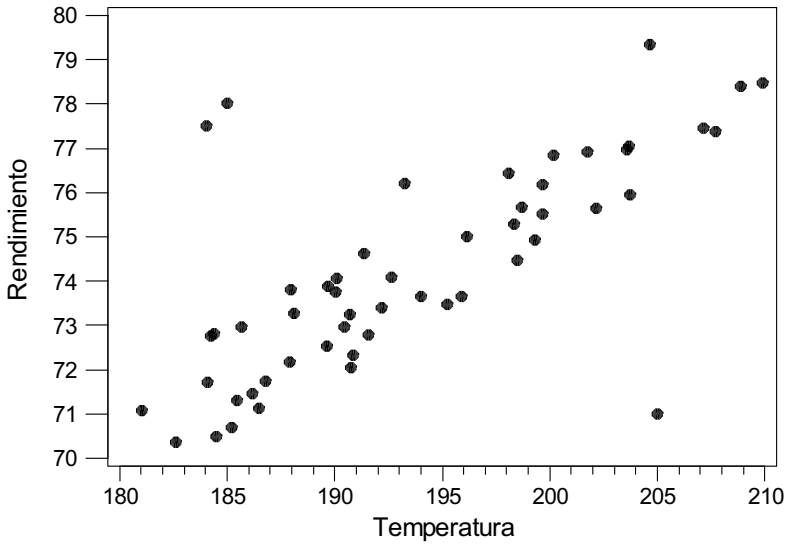


Figura 9.3. *Rendimiento obtenido al realizar una reacción química en función de la temperatura a la que se ha realizado*

En definitiva, ¿qué hacer ante una anomalía? La respuesta es: intentar averiguar el por qué se ha producido. Si está claro que la causa es un error, se elimina el valor y asunto resuelto. Si no es un error habrá que valorar la conveniencia de incluirla en el estudio, según sea la razón por la que se ha producido, la frecuencia con que se esperan valores similares, etc.

La verdad es que en algunos casos uno no sabe si mantener el valor atípico o quitarlo. Cuando se da esta situación una buena idea es hacer el análisis con y sin la presunta anomalía, y si se obtienen las mismas conclusiones la disyuntiva deja de tener importancia. En caso contrario quizá se puede salir de dudas recogiendo más datos, o también pueden aplicarse técnicas específicas de análisis en presencia de anomalías, sobre las que existen libros enteros como el de V. Barnett y T. Lewis: *Outliers in Statistical Data* Wiley, 1994.

10 ¿Qué es la curtosis (o *kurtosis*) y para qué sirve?

La curtosis es una medida de las llamadas "de forma" que cuantifica lo esbelta o aplanada que resulta una distribución de probabilidad (versión poblacional) o su equivalente cuando se refiere a un conjunto de datos (versión muestral). Se toma como referencia el valor que corresponde a la distribución Normal. Si una distribución tiene una curtosis mayor que la Normal hay que interpretarlo como que su parte central es más picuda (con más "apuntamiento") que una Normal con su misma desviación tipo, y si el valor es menor será más plana, lo cual se traduce en que sus colas son más "pesadas", es decir, que es más probable encontrar valores alejados de la media.

Sobre cuál es el valor de la curtosis para la distribución Normal existen 2 criterios. Su fórmula definida de forma natural es $[E(X-\mu)^4] / \sigma^4$ y para la Normal, con independencia del valor de sus parámetros, da un valor igual a 3. Pero para tomar este valor como referencia, también se define como la expresión anterior menos 3, de forma que para la Normal da cero. Cuando se dan valores de la curtosis, no siempre está claro cuál es el criterio con el que se ha calculado. Algunos textos se refieren a la curtosis como el resultado de aplicar la expresión anterior y definen el exceso de curtosis como ese valor menos 3.

A pesar de que la curtosis está relacionada con la forma de la distribución, no es una medida de variabilidad. Por ejemplo, una distribución uniforme definida en el intervalo $(-\sqrt{3}; \sqrt{3})$ tiene $\mu = 0$ y $\sigma = 1$, al igual que una N(0; 1), pero la uniforme tiene una curtosis de 9/5, mientras que para la Normal es igual a 3. Si comparamos distribuciones con forma de campana, en la Figura 10.1 tenemos una Normal con media cero y varianza 5/3, y una t de Student con 5 grados de libertad. Ambas tienen la misma media y la misma varianza, pero la curtosis de la t de Student es 9, mientras que para la Normal es, como siempre, igual a 3.

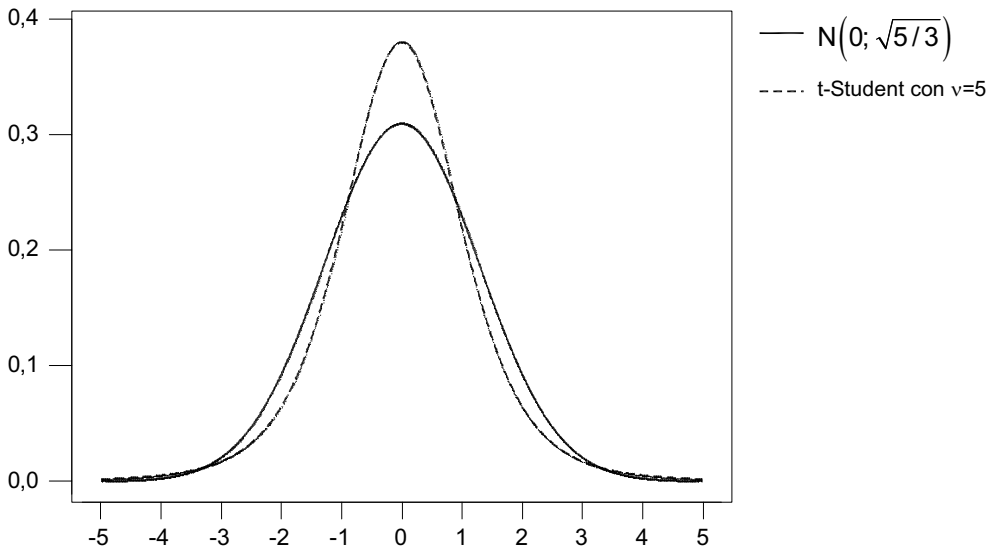


Figura 10.1. *Distribución t-Student con 5 grados de libertad (a trazos) y Normal con $\mu=0$ y $\sigma^2=5/3$ (trazo continuo). Ambas tienen la misma media y varianza pero distinta curtosis*

En la mayoría de paquetes estadísticos y en otros como Excel, cuando se piden las estadísticas descriptivas, además de las típicas medidas de tendencia central (media, mediana, moda) y de dispersión (varianza, desviación estándar, rango) se obtiene la curtosis y su inseparable compañero, el coeficiente de asimetría (en inglés *skewness*), que seguramente son las medidas menos atendidas del listado.

¿Para qué sirven? Ambas son útiles para caracterizar las distribuciones de probabilidad a nivel teórico. La Tabla 10.1 muestra los valores que presentan algunas de ellas.

Tabla 10.1. *Algunas distribuciones de probabilidad con sus valores de curtosis y coeficiente de asimetría*

Distribución	Curtosis	Coef. de asimetría
Uniforme	9/5	0
Normal	3	0
t-Student con ν grados de libertad	$3+6/(\nu-4)$, con $\nu\geq 5$	0, con $\nu\geq 4$
Chi-cuadrado con ν grados de libertad	$3+12/\nu$	$2\sqrt{2/\nu}$
Exponencial	9	2

También juegan un papel importante en algunas situaciones en las que para evaluar por simulación la robustez de un método o de un estimador, conviene probar con distribuciones de colas livianas y de colas pesadas (es decir, con distinta curtosis).

Sin embargo, tienen poco interés para describir la forma que presentan los valores de una muestra. Es mejor hacer un gráfico, como un diagrama de puntos o un histograma. El gráfico no puede ser sustituido por estas medidas, que tampoco aportan nada relevante cuando ya se tiene.

Además, son malos estimadores de los valores correspondientes a la población. La Figura 10.2 muestra los valores de exceso de curtosis obtenidos por simulación de

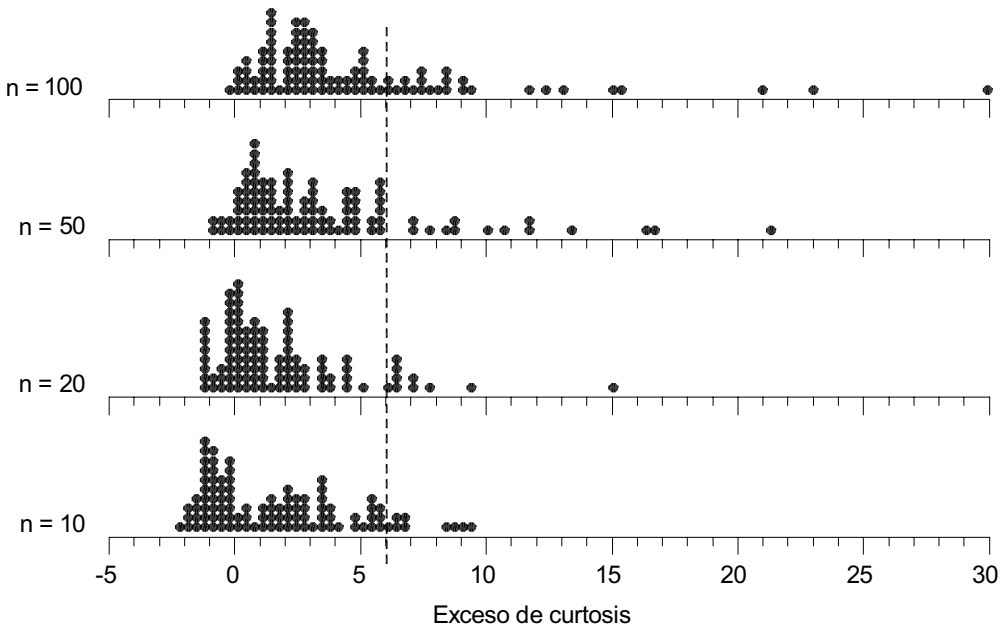


Figura 10.2. *Valores del exceso de curtosis (así lo da Minitab) obtenidos por simulación de muestras de tamaño 10, 20 50 y 100, de una población exponencial con λ=1. El valor para la población es igual a 6*

muestras de tamaño 10, 20 50 y 100, de una población exponencial con $\lambda=1$, a la que corresponde un exceso de curtosis igual a 6. Ya se ve que la estimación es sesgada, especialmente con muestras pequeñas, pero incluso con muestras tan grandes como n=100, un porcentaje alto de las estimaciones subestiman la curtosis verdadera. También se observa mucha variabilidad en los resultados, ya que los valores extremos afectan mucho al aparecer en la fórmula elevados a la cuarta potencia.

Si lo que pretendemos es utilizar los valores de la curtosis y el coeficiente de asimetría para intentar predecir el tipo de distribución de que provienen los datos, también lo tenemos mal. La Figura 10.3 muestra los valores de ambas medidas calculadas para 100 muestras de tamaño n = 50 de la misma distribución exponencial que hemos usado antes. Los verdaderos valores de la distribución (Exceso de curtosis = 6; Asimetría = 2) están marcados con una cruz. Ya se ve que los valores muestrales no permiten identificar este punto como aquel que representa los parámetros que se están estimando.

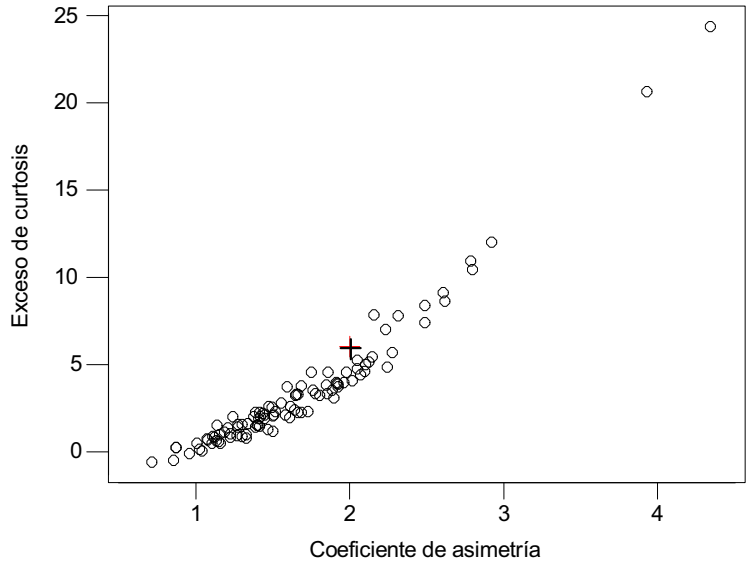


Figura 10.3. *Diagrama bivariante de los valores de exceso de curtosis y coeficiente de asimetría correspondientes a 100 muestras de tamaño 50 de una distribución exponencial. Los valores que corresponden a la población están marcados con una cruz*

En resumen, tanto la curtosis como el coeficiente de asimetría forman parte de las señas de identidad de una distribución de probabilidad y tiene interés estudiarlas en el marco de la caracterización de los modelos teóricos, pero como medidas descriptivas son muy poco fiables.

Distribuciones de probabilidad

11 ¿Cómo se sabe que una variable aleatoria concreta sigue una determinada distribución de probabilidad?

Seguramente las dudas no se presentan al elegir las distribuciones que podríamos llamar "instrumentales", es decir, aquellas que se utilizan básicamente como distribuciones de referencia. Sabemos que si $X \sim N(\mu; \sigma)$, entonces $(X-\mu)/\sigma$ sigue una distribución $N(0;1)$, y si estimamos σ a través de la desviación tipo de la muestra S, entonces $(X-\mu)/S$ sigue una *t de Student.* También sabemos que si n es el tamaño de una muestra aleatoria tomada de la población Normal anterior, entonces $S^2(n-1)/\sigma^2$ sigue una distribución *chi cuadrado.* Y si S_1^2 y S_2^2 son las varianzas de 2 muestras aleatorias de sendas poblaciones Normales e independientes con igual varianza, S_1^2/S_2^2 sigue una distribución *F de Snedecor.*

Otra cosa es cuando nos enfrentamos a una característica que varía de forma aleatoria y no sabemos a qué distribución se puede acudir para modelar su comportamiento. A continuación comentamos algunas ideas que pueden resultar útiles en este caso.

Si la variable es CONTINUA (puede tomar cualquier valor en un intervalo), presenta una distribución simétrica en torno a su valor central, y a medida que nos alejamos de la media disminuye la probabilidad de obtener valores, seguramente su comportamiento se puede modelar a través de una distribución *Normal.* Este es el caso de *la altura de las personas adultas de un mismo sexo y origen racial, el error experimental al medir una determinada magnitud, el peso de productos envasados o el grosor de una moneda, medido con un aparato de precisión.* Si todos los valores son igualmente probables, entonces la distribución se llama *uniforme.* Este es el caso, por ejemplo, de los *números aleatorios que dan algunas calculadoras*, que pertenecen a una distribución uniforme entre 0 y 1.

Otras variables continuas tienen una distribución no simétrica. Esta circunstancia se da cuando los valores se alejan con más libertad hacia un lado que hacia el otro, en el que normalmente hay un frontera (puede ser lo que llamamos "cero natural") que impide el paso. Por ejemplo, *la planitud de una pieza* puede medirse como la diferencia entre sus cotas verticales máxima y mínima cuando la pieza está horizontal. El valor mínimo para la planitud es 0, pero no hay límite para los valores grandes, con lo que si los valores están próximos al cero se obtiene una distribución asimétrica. En alguna de estas situaciones el logaritmo de los datos sigue una distribución que puede considerarse Normal, por lo que esta distribución se llama *lognormal.*

El tiempo entre 2 llegadas consecutivas a un servicio (*dos coches a una gasolinera, dos clientes a la caja de un supermercado*, ...) o en un contexto totalmente distinto, *el tiempo entre la emisión de 2 partículas en una sustancia radioactiva*, bajo ciertas condiciones se puede considerar que sigue una distribución *exponencial.* Esta distribución también es razonable en algunos casos para modelar el tiempo de vida, aunque en estos contextos de la fiabilidad la distribución estrella por su versatilidad es la de *Weibull.*

Si la variable es DISCRETA (solo pueden tomar valores a saltos, por ejemplo: 0, 1, 2, 3, ...) el caso más sencillo es cuando todos los valores tienen la misma probabilidad de aparecer (ejemplo *resultado de tirar un dado*). En este caso la distribución se llama *uniforme discreta.* Otro caso muy típico es el siguiente:

- Se realizan n pruebas independientes (el resultado de una no está afectado por el resultado de las anteriores).
- Cada prueba puede tener solo 2 resultados, que para entendernos designaremos como "éxito" y "fracaso".
- La probabilidad de éxito es p y, por tanto, la de fracaso es 1-p. Ambas probabilidades se mantienen constantes a lo largo de todas las pruebas.

En este caso, el número de éxitos X al realizar n experimentos es una variable aleatoria que sigue una distribución *binomial.* Ejemplos de variables que se ajustan a este modelo son el *número de caras que se obtienen al lanzar 10 veces una moneda al aire o el número de piezas defectuosas en una muestra aleatoria de 1.000*, sabiendo que la probabilidad de que una sea defectuosa es del 1%.

En ciertas situaciones, como cuando se considera el número de averías anuales que tiene una máquina, se podría aplicar el modelo binomial entendiendo que un año tiene 365 días y considerando que la probabilidad de que un día se estropee es, por ejemplo, 0,01 (modelo binomial con n=365 y p=0,01). Pero también se podría considerar que el año tiene 365×24 horas y la probabilidad de que se estropee en una hora es 0,01/24. O también se podrían establecer periodos de segundos, o milisegundos,... En estas situaciones en las que n se puede hacer crecer tanto como se quiera, disminuyendo p de forma que el producto np se mantiene constante, se dice que la variable sigue un modelo de *Poisson*. Situaciones en que es aplicable este modelo, bajo ciertas suposiciones, pueden ser, además del *número de averías mensuales de una máquina*, *el número de llamadas que se reciben en la centralita de una gran empresa cada 10 minutos.*

El disponer de un "catálogo" de distribuciones permite que cuando nuestra variable encaja en uno de los modelos ya descritos, no hace falta que deduzcamos las fórmulas para calcular sus probabilidades, ni su esperanza matemática u otras características de interés. Si se tienen datos también se puede realizar una prueba de ajuste para contrastar la hipótesis nula de que la distribución que se supone es la adecuada, aunque hay que tener en cuenta que si los datos son pocos va a ser difícil rechazar la hipótesis nula, cualquiera que esta sea.

En los libros también existen distribuciones "para hacer ejercicios", que solo son expresiones matemáticas en las que no se comenta cuál es el fenómeno a que se refieren ni cuál es el sentido físico de la variable en cuestión. Hay que entender estas distribuciones como instrumentos para practicar las propiedades de las distribuciones de probabilidad, aunque también es verdad que determinadas variables se pueden modelar con funciones específicas, "no catalogadas" como las que aparecen en ese tipo de ejercicios.

En cualquier caso, catalogado o no, no hay que confundir el modelo con la realidad. Uno de los ejemplos más socorridos es el de la altura de las personas para ilustrar la

distribución Normal, pero si tuviéramos las alturas exactas de los millones de habitantes adultos del planeta, podríamos comprobar que no se ajustan 'exactamente' a la conocida campana de Gauss, y tampoco lo harían si estratificamos por sexo, raza, o lo que sea. Se trata, como en los otros casos, de un buen modelo de referencia que permite realizar, seguramente con toda la precisión necesaria, estimaciones sobre la distribución de las alturas, pero no deja de ser un modelo teórico que no coincide exactamente con la realidad. Lo mismo ocurre con las otras distribuciones en las que, seguramente porque en la práctica no se cumplen exactamente las hipótesis que se consideran, no dejan de ser modelos teóricos (lo de teórico para un modelo es un calificativo innecesario) pero, eso si, enormemente útiles.

12 La media de una muestra es un número concreto. ¿Por qué se dice entonces que es una variable aleatoria?

Antes de hablar de medias hablaremos de observaciones individuales y utilizaremos el ejemplo de la distribución de las alturas. Empezaremos diciendo que la altura de una persona concreta es un número. Por ejemplo, Juan mide 1,73 metros, Antonio 1,82 y María 1,76. Estos valores son números fijos y concretos, puesto que Juan, al igual que Antonio y María, siempre miden lo mismo.

Otra cosa es si nos referimos a la altura de una persona genérica e indeterminada. La que en este momento puede estar pasando por delante de la puerta de su casa. ¿Qué altura tiene? No lo sabemos. La altura de una persona, así, a nivel general, es una variable aleatoria, que podemos modelar bastante bien a través de una distribución Normal.

Algo análogo ocurre con las medias de las muestras. La media de una muestra formada por unos individuos concretos es un número. Por ejemplo, si la variable que medimos son las alturas y la muestra está formada por Juan, Antonio y María, la media de esta muestra es 1,77. Pero si hablamos de la muestra de 3 individuos tomados al azar, la media de esa muestra es una variable aleatoria, ya que está formada por observaciones individuales, que a su vez también son variables aleatorias.

Lo más interesante de este tema es que la media muestral se distribuye siempre con la misma media que las observaciones individuales, con una varianza que es la enésima parte (siendo n el tamaño de la muestra) de la que tiene esa distribución y además, muy frecuentemente, su distribución es muy próxima a la Normal[1].

Pensemos, por ejemplo, en la altura de 20 personas, si representamos sus valores en un diagrama de puntos, podemos obtener un gráfico como el de la Figura 12.1.

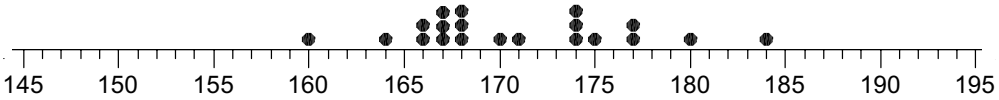


Figura 12.1. *Diagrama de puntos de las alturas de un grupo de 20 personas*

Sin embargo, si pensamos en la altura media de grupos de 25 personas, ya no podemos dar valores tan extremos, puesto que aunque es perfectamente posible que una persona

[1] Si la población es Normal, la media muestral sigue una distribución que también es Normal. Si la población no es Normal, hace falta un cierto tamaño de muestra para poder considerar que la distribución de la media muestral sea Normal. Dicho tamaño de muestra depende de lo distinta de la Normal que sea la población. Para poblaciones con ligeras diferencias respecto a la Normal (ligeramente sesgada hacia un lado, por ejemplo), no hay que preocuparse demasiado y basta con muestras de 3 o 4 observaciones. Si la población es muy distinta de la Normal es necesario que la muestra tenga un cierto tamaño para que la aproximación Normal funcione bien a efectos prácticos.

mida 1,85 metros, es prácticamente imposible que esta sea la altura media de un grupo de 25 personas tomadas al azar, ya que en este grupo cabe esperar que haya un número parecido de personas por encima y por debajo de la media general, de forma que los valores de sus alturas se compensarán, dando un valor medio cercano a la media de la población. En la Figura 12.2 se muestra el diagrama de puntos de unos valores que podrían corresponder a las alturas medias de 20 grupos con 25 individuos cada uno. Estos valores se han obtenido por simulación suponiendo que las alturas individuales siguen una N(1,70 m; 0,07 m). Obsérvese como su dispersión es mucho menor que para las alturas individuales

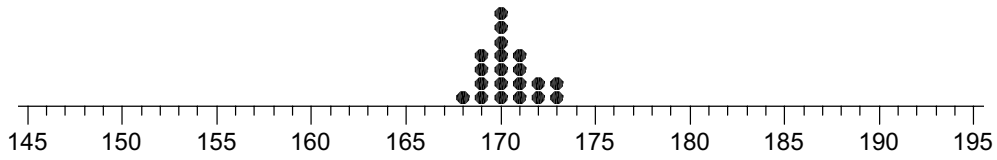


Figura 12.2. *Valores que podrían corresponder a las alturas medias de 20 grupos con 25 individuos cada uno*

Las distribuciones de probabilidad a que pertenecen ambos conjuntos de datos están representadas en la Figura 12.3. La distribución de las alturas individuales la hemos supuesto Normal y le hemos atribuido unos parámetros que nos han parecido razonables (μ=1,70 m y σ= 0,07 m). Dando como buena la distribución de las alturas individuales, podemos afirmar que las medias seguirán también una distribución Normal, con la misma media que la población, y una desviación tipo que será: $\sigma_{\bar{X}} = \sigma / \sqrt{n}$, en nuestro caso: $0,07 / \sqrt{25} = 0,014$.

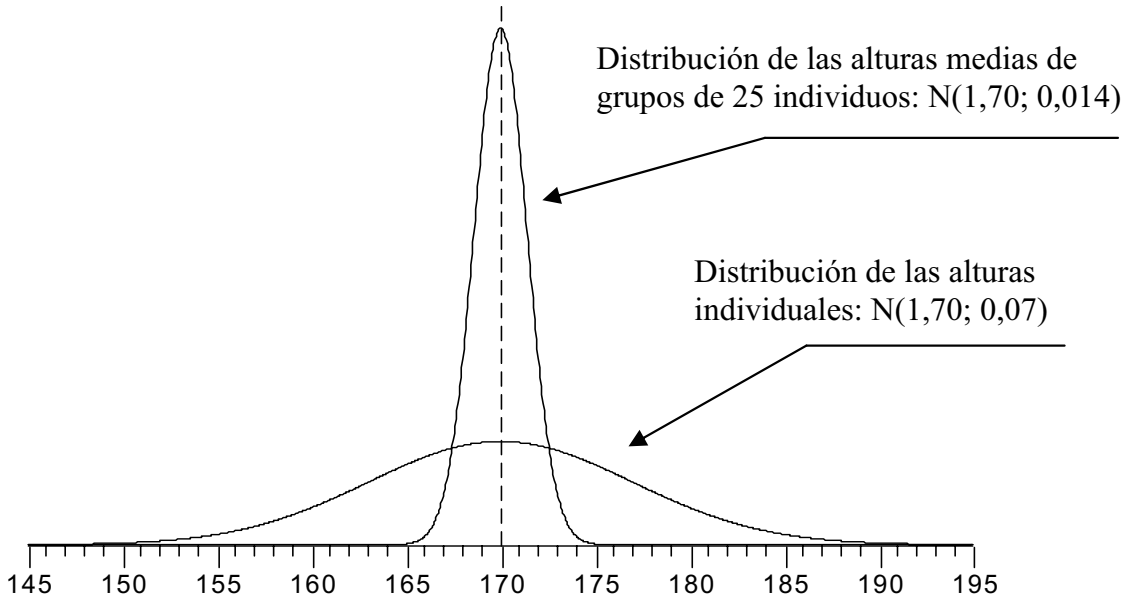


Figura 12.3. *Distribución que hemos supuesto para la altura de las personas, N(1,70 m; 0,07 m) (la más ancha) y la que se deduce de esta para las medias de grupos de 25 individuos (la mas esbelta)*

En definitiva, la media muestral, protagonista destacada en muchos estudios estadísticos, es una variable aleatoria nada misteriosa, con un comportamiento noble y fácil de prever.

13 ¿Por qué la función densidad de probabilidad de la distribución Normal es la que es?

Se trata de la función que define la forma de campana de la distribución Normal:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Fue obtenida por primera vez por Abraham de Moivre en 1733 como distribución límite de la distribución binomial. Posteriormente, K. F. Gauss (1777-1855) descubrió muchas de sus propiedades y la popularizó. Su demostración más general es la del llamado "Teorema Central del Límite", y esta no llegó, de forma completa y rigurosa, hasta bien entrado el siglo XX.

Pero no vamos a describir aquí ninguna de estas demostraciones, ya que su seguimiento requiere de unos conocimientos y habilidades matemáticas, que no son frecuentes ni necesarias en aquellos que solo desean entender y sacar partido de las técnicas estadísticas. Lo que sí podemos hacer, utilizando matemáticas sencillas y consideraciones geométricas, es justificar el porqué la fórmula es como es.

La Figura 13.1 muestra el histograma correspondiente a 10.000 valores generados aleatoriamente de una distribución Normal con media μ=10 y desviación tipo σ=2. El perfil suavizado de este histograma será una buena aproximación a la distribución teórica.

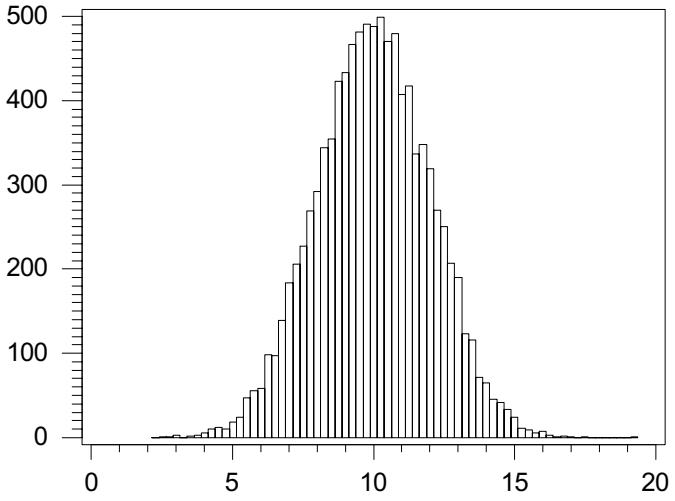


Figura 13.1. *Histograma correspondiente a 10.000 valores de una N(10; 2)*

La primera observación que puede hacerse es que la forma de caída de los lados del histograma recuerda a la función exponencial. La parte de la izquierda, en la que la función es creciente, podría ser de la forma $f(x) = e^x$, pero para poder representar la parte derecha de forma simétrica a esta, utilizaremos la función $f(x) = e^{x-10}$ para la

parte creciente de la izquierda $f(x)=e^{10-x}$ para la parte de la derecha. Hemos representando ambas funciones en la Figura 13.2.

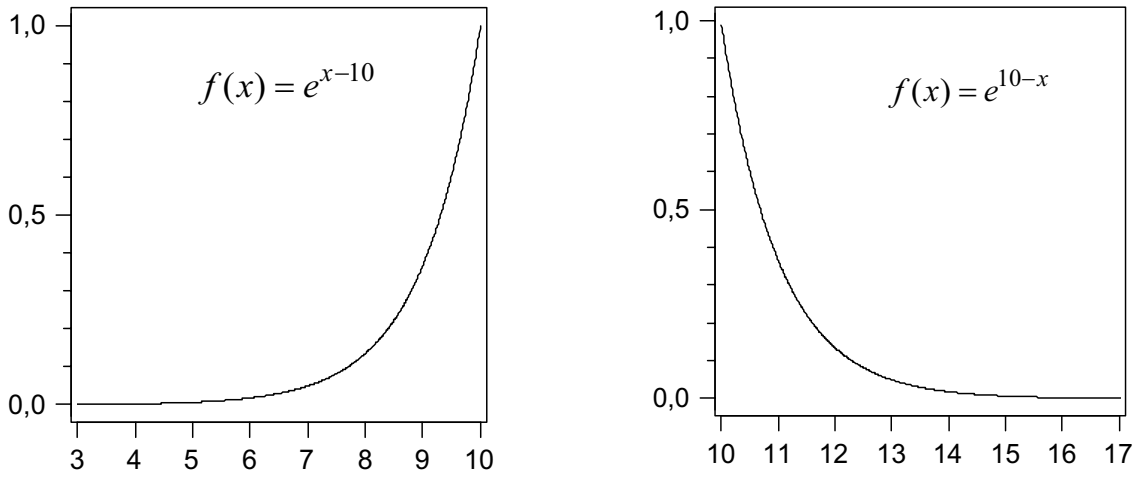


Figura 13.2. *Representación gráfica de las funciones exponenciales que se indican*

Ahora debemos colocar en una sola función la parte creciente y la decreciente. Esto se consigue mediante la expresión $f(x)=e^{-|x-10|}$ (Figura 13.3).

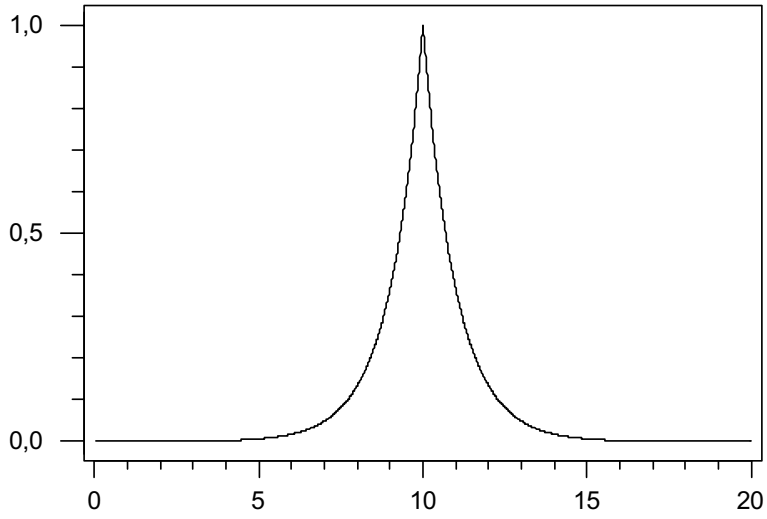


Figura 13.3. *Representación gráfica de la función* $f(x)=e^{-|x-10|}$

Ya tenemos una función que tiene una forma parecida a la Normal en las colas pero que presenta un pico en su máximo. Este pico, punto en el que la función no tiene derivada, es uno de los malos comportamientos característicos de la función valor absoluto. Por tanto, si el problema parece estar en la expresión $|x-10|$, es razonable pensar en sustituirla por el cuadrado de la diferencia, $(x-10)^2$.

La Figura 13.4 muestra la función $f(x)=e^{-(x-10)^2}$ superpuesta con el histograma de los datos. Vamos por buen camino, efectivamente el pico ha desaparecido y ya tenemos forma de campana, aunque ahora se trata de ensancharla.

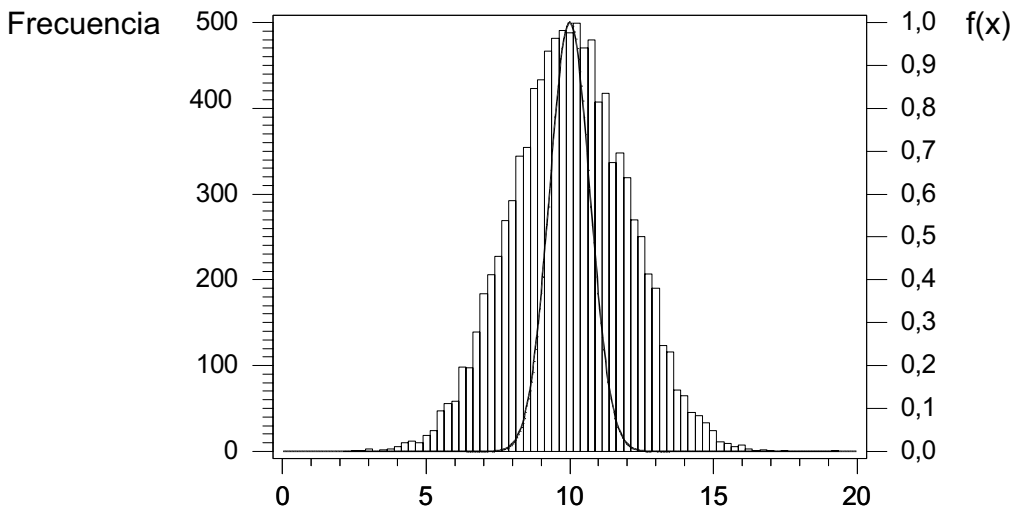


Figura 13.4. *Función $e^{-(x-10)^2}$ superpuesta al histograma de los datos. La escala horizontal es común, la vertical de la izquierda corresponde al histograma, y la derecha a la función*

Ensanchar significa aumentar el valor de las ordenadas para $x \neq 10$ manteniendo la forma de la función, y esto se consigue dividiendo el exponente por una constante k, más concretamente, dividiendo por $2\sigma^2$.

Para comprobarlo vamos identificar cuáles son –aproximadamente– las ordenadas de la función que se ajusta al perfil del histograma, para los valores de x correspondientes a μ (=10), $\mu\pm\sigma$ (8 y 12), $\mu\pm2\sigma$ (6 y 14), $\mu\pm3\sigma$ (4 y 16).

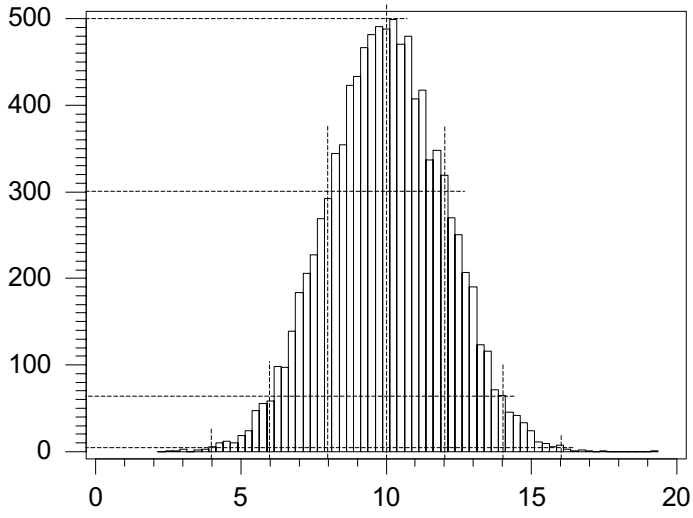


Figura 13.5. *Ordenadas aproximadas de la función que se ajusta al perfil del histograma, para los valores de x correspondientes a μ; $\mu\pm\sigma$; $\mu\pm2\sigma$; $\mu\pm3\sigma$*

Los valores obtenidos se encuentran en la Tabla 13.1. Podemos observar que en la función que andamos buscando, la ordenada en $\mu \pm \sigma$ debe ser aproximadamente el 60% de la ordenada máxima. En $\mu \pm 2\sigma$ debe ser del orden del 13%, y en $\mu \pm 3\sigma$ del orden del 1%. En la misma tabla se tienen los valores obtenidos con la función que estamos probando, dividiendo el exponente por $2\sigma^2$. Está claro que da muy buen resultado.

Tabla 13.1. *Ordenadas en el histograma de los datos y en la función* $f(x) = e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Valores de x	Ordenadas (aproximadas) en el histograma	Proporción respecto al máximo	$f(x) = e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
μ: 10	500	500/500 = 1	1
$\mu \pm \sigma$: 8 y 12	300	300/500 = 0,6	0,607
$\mu \pm 2\sigma$: 6 y 14	65	65/500 = 0,13	0,135
$\mu \pm 3\sigma$: 4 y 16	5	5/500 = 0,01	0,011

La Figura 13.6 muestra la superposición de la función $f(x) = e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, en la que –recordemos– μ= 10 y σ=2, con el histograma de los datos. El ajuste es excelente.

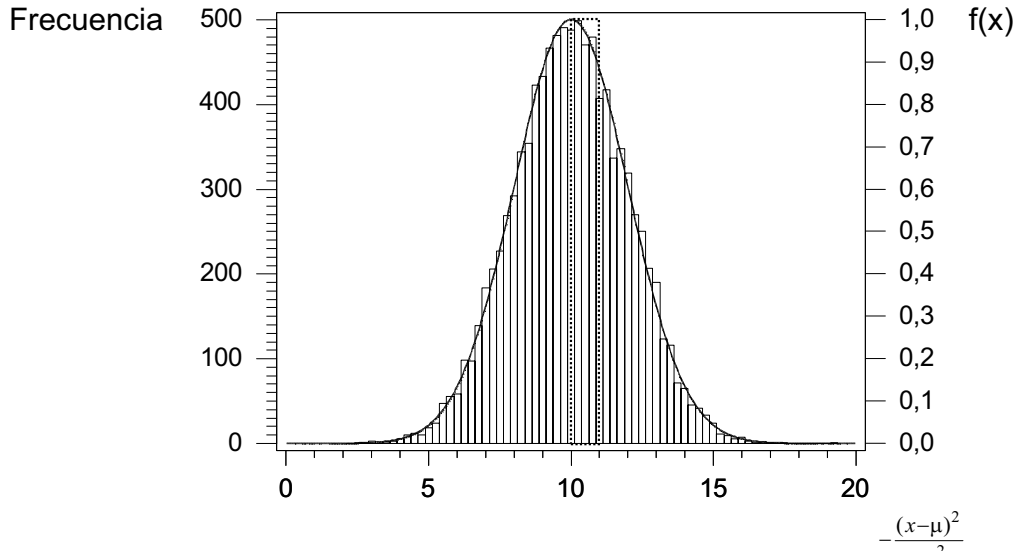


Figura 13.6. *Superposición del histograma de los datos y la función* $f(x) = e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ *, con* μ*=10 y* σ*=2*

Pero esta no es la función densidad de probabilidad de la distribución Normal, no es ni siquiera una función densidad de probabilidad. Para serlo es necesario que el área definida por la curva sea igual a 1, y en nuestra función esto no es así. Puede observar que, por ejemplo, la porción de campana situada entre los valores de x=10 y x=11, (recuadro punteado en la Figura 13.6) ya tiene un área casi igual a 1.

Si el área es igual a K, entonces $\frac{1}{K} f(x)$ mantendrá la misma forma y con las mismas proporciones, ya que solo cambia un factor de escala en el eje de ordenadas, y ahora el área sí será igual a 1.

El valor de K es $\sigma\sqrt{2\pi}$. Su deducción precisa el cálculo de una integral que no es inmediata, pero que se encuentra en muchos libros[1]. En nuestro ejemplo, en el que σ=2, puede comprobarlo calculando el área (la integral) con la ayuda de un programa de

[1] Por ejemplo, en *Probabilidad y Estadística* de M. H. DeGroot. Addison-Wesley Iberoamericana, 1988.

cálculo simbólico como el de la Figura 13.7 . Si lo hace obtendrá 5,013, que es exactamente $2\sqrt{2\pi}$.

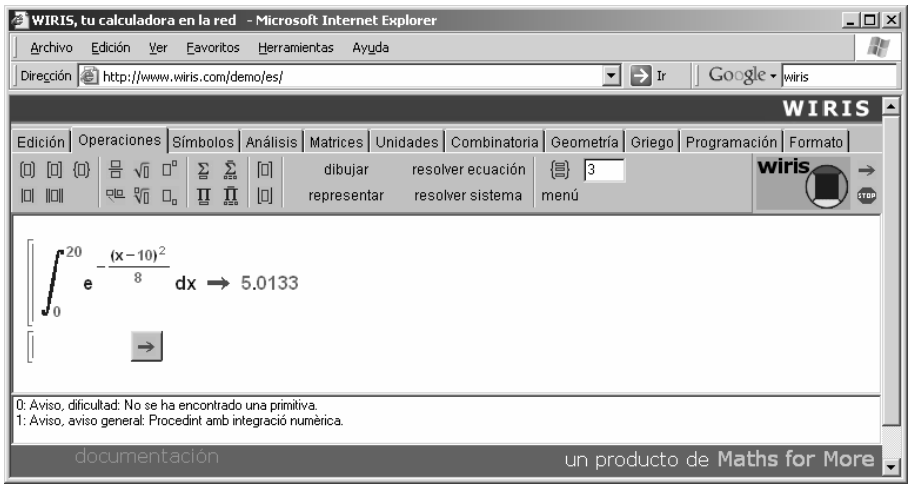


Figura 13.7. *Cálculo de la integral con un programa al que se puede acceder a través de internet. El denominador del exponente es $2\sigma^2$, con $\sigma=2$*

14 ¿Por qué las probabilidades calculadas a través de la Normal estandarizada coinciden con las buscadas en la distribución de interés?

Lo veremos primero a través de un enfoque gráfico que no precisa conocimientos de cálculo. Sea $X \sim N(\mu; \sigma)$ y deseamos calcular $P(X > x)$. Si definimos $Y = X - \mu$, está claro que $Y \sim N(0; \sigma)$ ya que $E(X-\mu) = E(X) - \mu = 0$, y $V(X-\mu) = V(X) = \sigma^2$.

Con ayuda de la Figura 14.1 es fácil observar que $P(X > x) = P(Y > x-\mu)$ ya que la forma de las 2 distribuciones es idéntica (tienen la misma σ) y al punto x en la distribución de X le corresponde el $x-\mu$ en la distribución de Y.

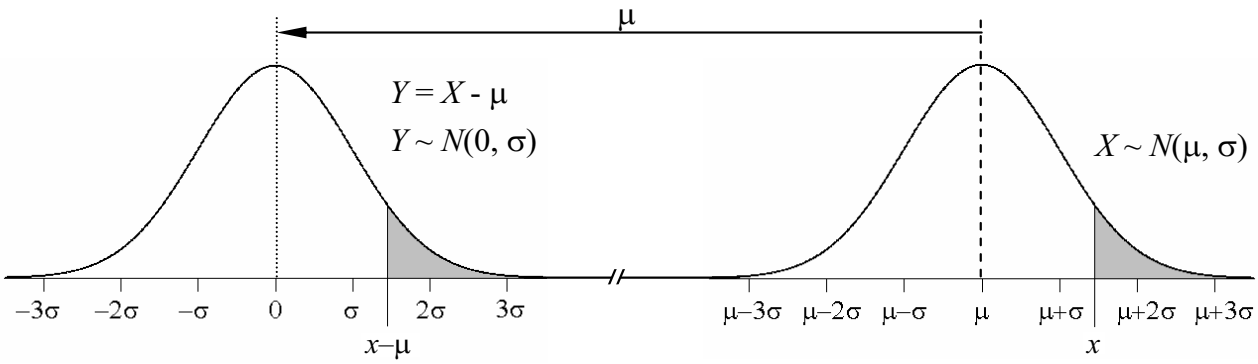


Figura 14.1. *Transformación de la distribución $X \sim N(\mu; \sigma)$ en la $Y \sim N(0; \sigma)$*

Si $Y \sim N(0; \sigma)$, entonces $kY \sim N(0; k\sigma)$, ya que $E(kY) = kE(Y) = 0$ y $V(kY) = k^2\sigma^2$. Veamos ahora cómo representar la distribución de kY.

Si representamos $f(Y)$ y $f(kY)$ utilizando los mismos ejes, las formas de las distribuciones serán distintas, por tener distinto valor de σ, pero también podemos representarlas usando la misma forma para la distribución (misma campana) y cambiando las escalas de los ejes, tal como se indica en la Figura 14.2.

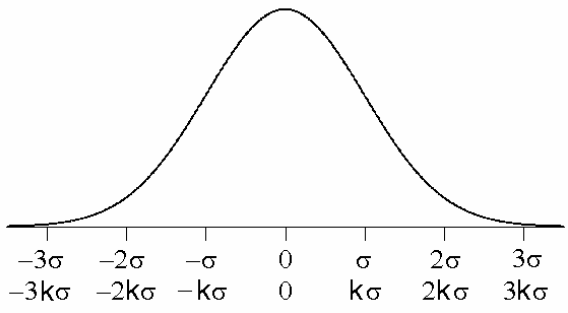


Figura 14.2. *Distribución que representa una $N(0;\sigma)$ si se considera la escala superior, y una $N(0; k\sigma)$ si se toma la escala inferior*

Recordando que la función densidad de probabilidad de $Y \sim N(0; \sigma)$ es:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{y^2}{2\sigma^2}}$$

se deduce que:

$$f(ky) = \frac{1}{k\sigma\sqrt{2\pi}} e^{-\frac{k^2y^2}{2k^2\sigma^2}} = \frac{1}{k} f(y)$$

Por tanto, las campanas pueden representarse de forma idéntica, multiplicando por k el eje de abcisas y dividiendo por el mismo valor el de ordenadas (que no hemos representado).

Si hacemos $k=1/\sigma$ tendremos $Z = Y/\sigma$, o lo que es lo mismo: $Z = (X - \mu)/\sigma$. Por lo comentado anteriormente, la distribución de Z se puede representar con la misma campana que la utilizada para la distribución de Y, tal como se indica en la Figura 14.3.

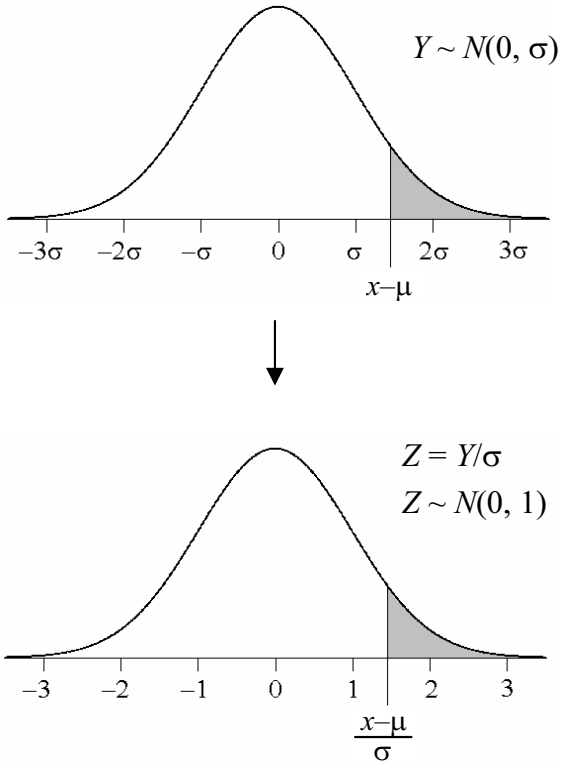


Figura 14.3. *Distribuciones de $Y \sim N(0; \sigma)$ y $Z \sim N(0; 1)$ representadas con la misma campana, pero con diferentes escalas en los ejes*

Si la forma de las campanas es la misma y solo hemos sustituido el punto $x–\mu$ por el $\frac{x-\mu}{\sigma}$ resulta claro que $P\left(Z > \frac{x-\mu}{\sigma}\right) = P(Y > X–\mu)$ y por tanto, $P(X > x) = P\left(Z > \frac{x-\mu}{\sigma}\right)$.

• • • • •

Vamos a verlo ahora con un razonamiento analítico utilizando conocimientos de cálculo. Sabemos que si $X \sim N(\mu; \sigma)$, entonces:

$$P(X \le x) = F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\, dx$$

Se trata de demostrar que esta probabilidad es exactamente igual a la que se calcula para una variable Z con distribución $N(0; 1)$ cambiando la "x" por $(x-\mu)/\sigma$. Es decir, hay que probar que:

$$P(X \le x) = F_X(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\, dx = F_Z\left(\frac{x-\mu}{\sigma}\right) = \int_{-\infty}^{\left(\frac{x-\mu}{\sigma}\right)} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2}\, dz$$

Para hacerlo partimos de la expresión de $F_X(x)$ y hacemos el cambio de variable $Z = \dfrac{X-\mu}{\sigma}$, lo cual implica que $X = \mu + Z\sigma$, y por tanto que $\dfrac{dx}{dz} = \sigma$, de donde: $dx = \sigma\, dz$.

Sustituyendo en la expresión de $F_X(x)$ queda:

$$F_Z\left(\frac{x-\mu}{\sigma}\right) = \int_{-\infty}^{\left(\frac{x-\mu}{\sigma}\right)} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{\mu+z\sigma-\mu}{\sigma}\right)^2} \sigma\, dz =$$

$$\int_{-\infty}^{\left(\frac{x-\mu}{\sigma}\right)} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz = P\left(Z \le \frac{x-\mu}{\sigma}\right)$$

Un posible error que da al traste con la demostración es sustituir algebraicamente los límites de la integral cambiando x por $\mu+z\sigma$. Lo que debe hacerse (y es lo que hemos hecho) es obtener el valor de Z, cuando $X = -\infty$ y el valor de Z, cuando $X = x$, es decir, de transformar los límites que estaban en términos de X, a limites en términos de Z.

15 Yo mido 1,68. ¿Por qué la probabilidad de que una estatura sea 1,68 calculada con la distribución Normal es 0?

Se trata de una aparente contradicción debido a que en la práctica tratamos como discretas a las variables que en teoría consideramos como continuas.

Cuando decimos que la variable aleatoria X sigue una distribución N(μ; σ), estamos diciendo también que X es una variable continua que puede tomar infinitos valores. Sabemos que la probabilidad de que tome valores comprendidos entre a y b es igual al área definida por su función densidad de probabilidad entre estos valores, es decir:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Por tanto:

$$P(X = 1,68) = P(1,68 \leq X \leq 1,68) = \int_{1,68}^{1,68} f(x)dx = 0$$

Pero cuando decimos que alguien mide 1,68 metros no queremos decir que mide exactamente 1,680000000… con infinitos ceros, sino que entendemos que el valor se ha redondeado, y lo que queremos decir es que su altura está comprendida entre 1,675 y 1,685 (si fuera menor de 1,675 lo habríamos aproximado a 1,67 y si mayor de 1,685 a 1,69)

Suponiendo que las alturas se distribuyan según una Normal con media μ=1,70 m y desviación tipo σ=0,07 m, la probabilidad de que una persona mida 1,68 m (entendido como valor redondeado, que es como hablamos), es de 0,0547. Tranquilo, usted existe.

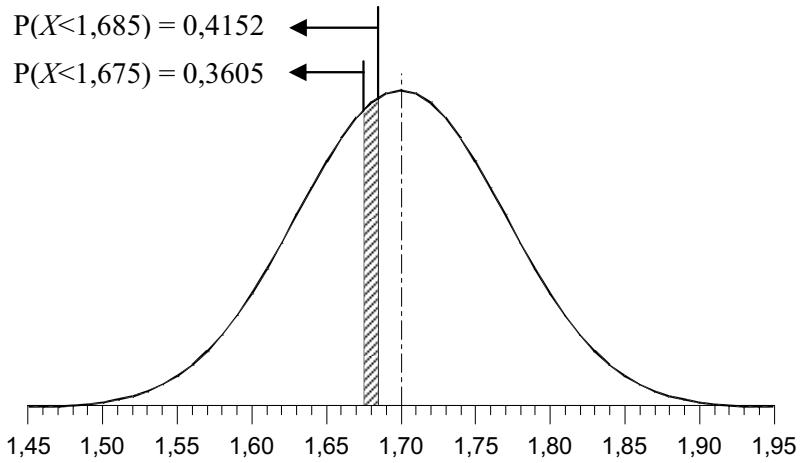


Figura 15.1. *Probabilidad de que X ~ N(1,70; 0,07) tome valores comprendidos entre 1,675 y 1,685*

Todas las variables continuas se discretizan para tratar con ellas en la práctica. Casi siempre se redondea, como en el caso de las alturas, aunque en algunos casos se trunca, como hacemos con las edades. Si hoy es día de las elecciones generales y usted cumple

los 18 años mañana, usted hoy tiene 17+364/365 = 17,997 años. Pero seguramente no le dejarán votar.

Observación final:

Un argumento falaz, pero muchas veces convincente haciendo mención a aquello de los casos favorables partido por los casos posibles, es considerar que si la variable puede tomar infinitos valores, la probabilidad de que tome uno en concreto es cero. Esto es falso porque la regla de casos favorables partido por casos posibles solo vale cuando todos los sucesos tienen la misma probabilidad de ocurrir, y este no es nuestro caso. Por otra parte, es perfectamente posible que una variable pueda tomar infinitos valores y que ninguno de ellos tenga probabilidad cero. Piense en la distribución de Poisson, por ejemplo.

16 ¿Existen variables aleatorias que presenten un comportamiento "contrario" a la distribución Normal, siendo los valores más probables los de los extremos?

Sí las hay, aunque son raras. Veamos, por curiosidad, algunos ejemplos:

Distribución de la mortalidad por edades

Si la variable "edad al morir" siguiera una distribución Normal, la mayoría de personas moriría en torno a los 40 años (siendo optimistas) y a medida que nos fuéramos alejando de esta edad, tanto por arriba como por abajo, habrían cada vez menos defunciones.

Pero sabemos que esto no es así. La distribución de la edad al morir tiene forma de U en los países poco desarrollados, con una alta mortalidad infantil, y forma de J en los países más desarrollados. Es decir, las frecuencias crecen hacia los extremos. La Figura 16.1 muestra el número de defunciones por edad y sexo (los hombres son las barras de la izquierda) en diferentes países. Los datos se han tomado de la página web de la organización mundial de la salud: *www3.who.int/whosis/menu.cfm* (para cada país se ha tomado el año más reciente en que se tienen datos).

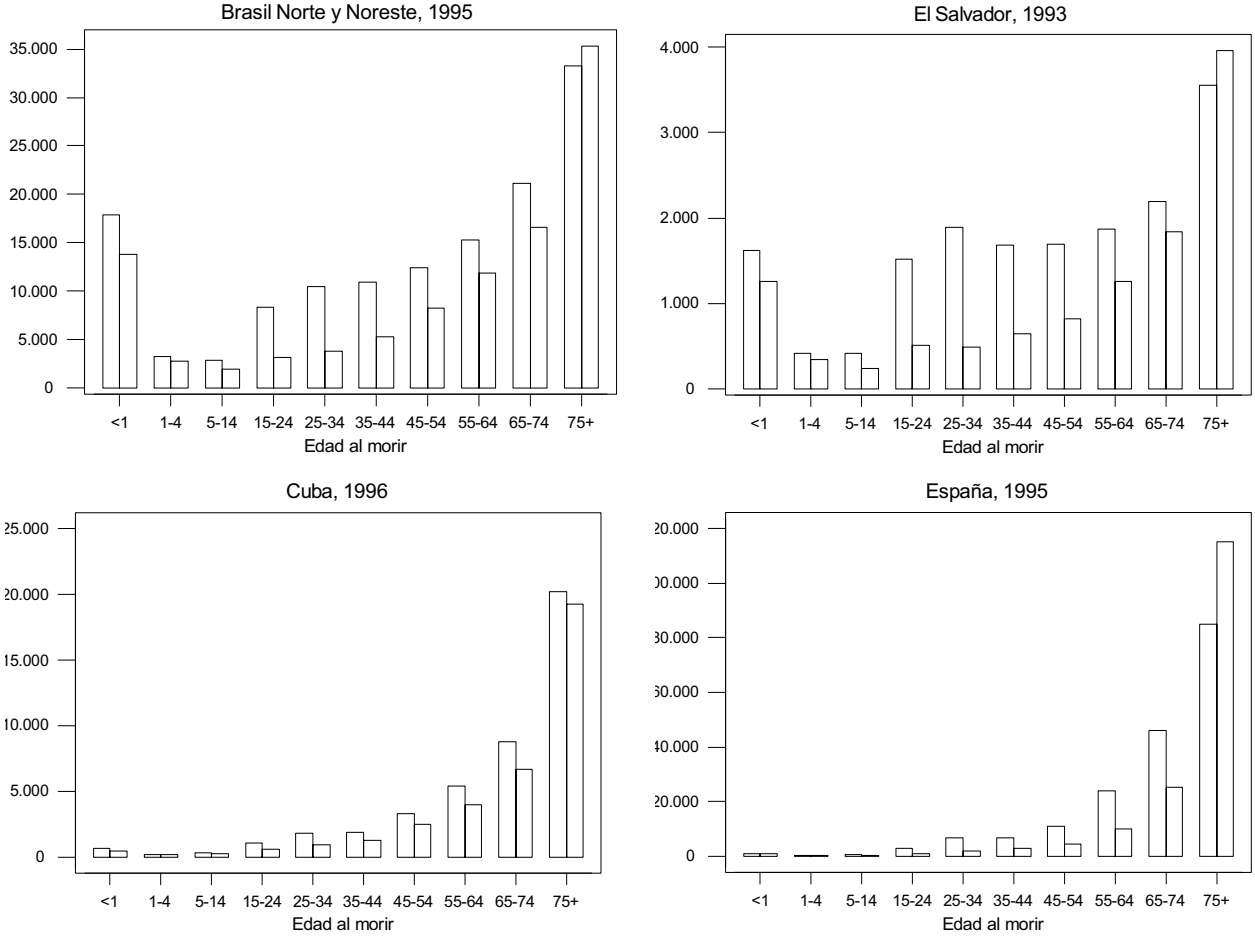


Figura 16.1. *Número de defunciones por edad y sexo (los hombres son la barra de la izquierda) en diferentes países*

Distribución de la nubosidad diaria

Algún texto indica que la distribución de la variable "porcentaje de cielo cubierto" tiene forma de U. Es decir, que son más frecuentes los días totalmente cubiertos, o totalmente despejados, que aquellos en los que el cielo está cubierto al 50%.

Para comprobar si esto es cierto, hemos analizado unos archivos que se pueden obtener en la página web de la Agencia de Protección Medioambiental de EE UU: *www.epa.gov/ceampubl/mmedia/metdata*. Estos archivos contienen datos meteorológicos de 237 estaciones durante el periodo 1961-1990, y entre los datos que incluye para cada estación está el de las décimas partes de cielo cubierto día a día.

Sin hacer un estudio exhaustivo, hemos visto que en algunos lugares la forma de U se cumple y en otros no. Por ejemplo, no se cumple en Nueva York ni en Los Ángeles, y sí se cumple en Madison y San Francisco. Ver Figura 16.2.

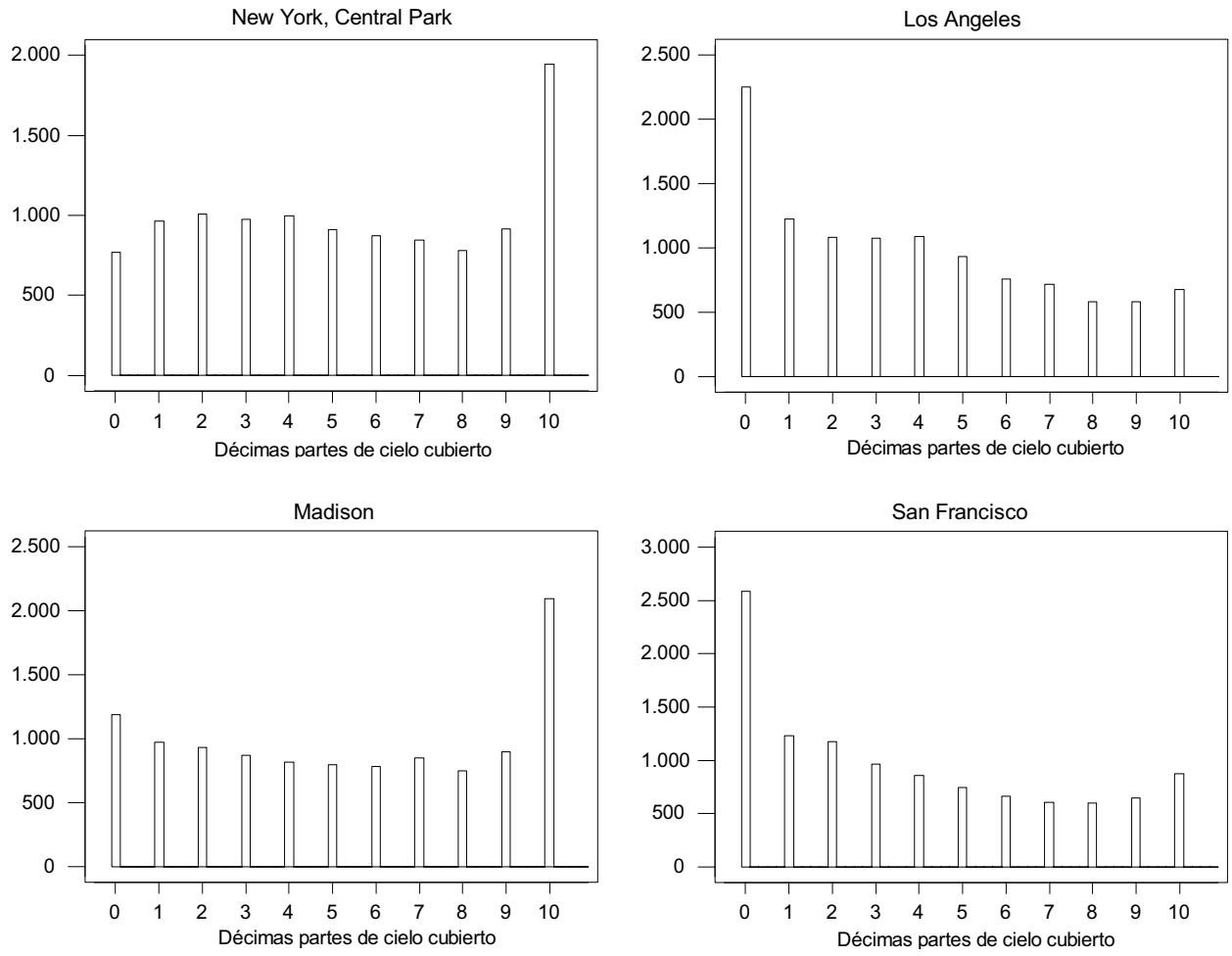


Figura 16.2. *Distribución de la variable "décimas partes de cielo cubierto" en mediciones diarias realizadas en estaciones meteorológicas situadas en las ciudades que se indican, en el periodo 1961-1990*

Distribución del coeficiente de correlación muestral cuando el coeficiente poblacional es ρ=0 y el tamaño de muestra es *n*=3.

Este es el caso que nos parece más curioso. Cuando ρ=0 la función densidad de probabilidad del coeficiente de correlación *r* para muestras de tamaño *n* viene dada por la expresión[1]:

[1] El símbolo ξ representa la función gamma de Euler. La deducción de esta fórmula puede verse en T. W. Anderson *Ann Introduction to Multivariate Statistical Analysis,* Wiley, 1994, donde también se deduce la expresión de *f*(r) cuando p×0.

$$f(r|\rho=0)=\frac{\Gamma\left[\frac{1}{2}(n-1)\right]}{\Gamma\left[\frac{1}{2}(n-2)\right]\sqrt{\pi}}\left(1-r^2\right)^{\frac{1}{2}(n-4)}$$

Sustituyendo por n=3 queda: $f(r|\rho=0;n=3)=\frac{1}{\pi\sqrt{1-r^2}}$

Y representándola, se tiene la forma:

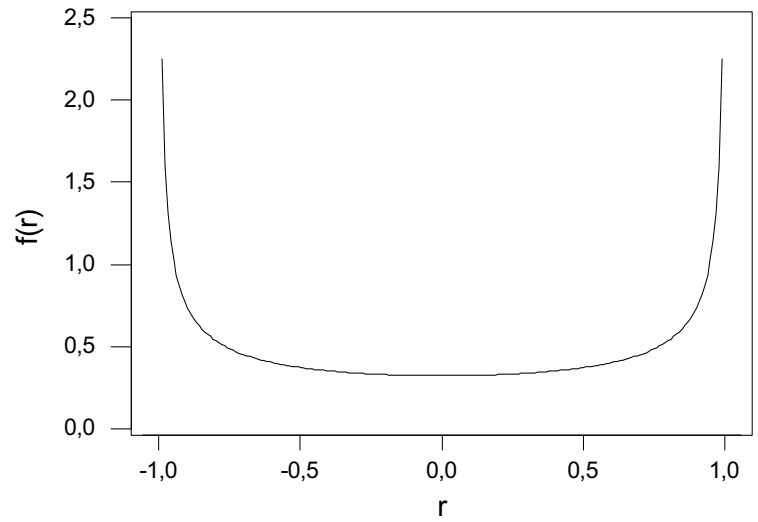


Figura 16.3. *Función densidad de probabilidad del coeficiente de correlación muestral cuando ρ=0 y el tamaño de muestra n=3*

Para constatar que esto es así hemos generado por simulación 10.000 pares de muestras aleatorias de una N(0; 1) con tamaño n=3. Para cada par (como ambas son aleatorias, son también independientes) hemos calculado su coeficiente de correlación, y el histograma de los valores obtenidos es el que se muestra en la Figura 16.4.

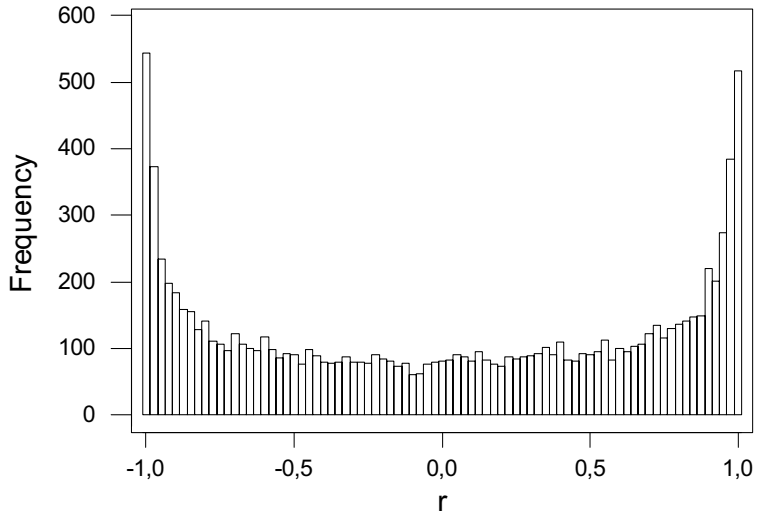


Figura 16.4. *Histograma de los coeficientes de correlación de 10.000 pares de muestras independientes con tamaño n=3, generadas aleatoriamente de una N(0; 1)*

Si n=4, es fácil comprobar que $f(r|\rho=0;n=4)=\Gamma(3/2)/\sqrt{\pi}$, y por tanto es constante, y como está definida entre -1 y 1 podemos asegurar que valdrá 0,5 sin necesidad de hacer cálculos.

Cuando n=5 la moda ya está en 0, y a medida que aumenta n va apareciendo la "inevitable" forma de campana.

17 ¿De dónde sale la fórmula de la distribución de Poisson?

Hay que reconocer que la fórmula de la distribución de Poisson, con el número *e* incluido, es una fórmula curiosa. Vamos a deducirla a partir de la distribución binomial cuando el número de pruebas *n* tiende a infinito y la probabilidad de éxito *p* tiende a cero, manteniéndose constante el valor *np,* al que llamamos λ.

Sabemos que si *X* sigue una distribución binomial con parámetros *n*, *p*:

$$P(X=x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

$$= \frac{n!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}$$

Se sustituye *p* por λ/*n*

$$= \frac{n(n-1)...(n-x+1)}{x!}\cdot\frac{\lambda^x}{n^x}\cdot\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}$$

Se simplifica la expresión del número combinatorio y se realizan cambios evidentes

$$= \underbrace{\frac{n(n-1)...(n-x+1)}{n^x}}_{1}\cdot\underbrace{\frac{\lambda^x}{x!}}_{2}\cdot\underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{3}\underbrace{\left(1-\frac{\lambda}{n}\right)^{-x}}_{4}$$

Se intercambian los valores de n^x y $x!$ en los denominadores de los 2 primeros términos

Dividimos la expresión en 4 partes que analizaremos por separado

Analizamos ahora cada una de las partes de la expresión anterior:

1. $$\frac{n(n-1)...(n-x+1)}{n^x} = \frac{n}{n}\cdot\frac{n-1}{n}\cdot\frac{n-2}{n}\cdot...\cdot\frac{n-x+1}{n} =$$

$$= 1\cdot\left(1-\frac{1}{n}\right)\cdot\left(1-\frac{2}{n}\right)\cdot...\cdot\left(1-\frac{x-1}{n}\right)$$

Está claro que si $n \to \infty$ manteniéndose x constante, cada uno de los términos tiende a 1 y, por tanto,

$$\lim_{n\to\infty}\frac{n(n-1)...(n-x+1)}{n^x} = 1$$

2. La expresión $\frac{\lambda^x}{x!}$ la dejaremos tal como está.

3. El análisis de esta tercera parte va a ser un poco más largo. Partiremos de la definición del número *e* como $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$ e intentaremos poner nuestra expresión de una forma parecida a esta.

Para empezar, puede comprobarse fácilmente que se verifica la igualdad:

$$\left(1-\frac{\lambda}{n}\right)^n=\left[\left(1-\frac{1}{n/\lambda}\right)^{\frac{-n}{\lambda}}\right]^{-\lambda}$$

y operando ahora solo con la expresión del interior del corchete, tenemos:

$$\left(1-\frac{1}{n/\lambda}\right)^{\frac{-n}{\lambda}}=\left(\frac{n/\lambda}{n/\lambda}-\frac{1}{n/\lambda}\right)^{\frac{-n}{\lambda}}=\left(\frac{n/\lambda-1}{n/\lambda}\right)^{\frac{-n}{\lambda}}=\left(\frac{n/\lambda}{n/\lambda-1}\right)^{\frac{n}{\lambda}}$$

y también podemos hacer:

$$\left(\frac{n/\lambda}{n/\lambda-1}\right)^{\frac{n}{\lambda}}=\left(1+\frac{1}{n/\lambda-1}\right)^{\frac{n}{\lambda}}=\left(1+\frac{1}{n/\lambda-1}\right)^{\frac{n}{\lambda}-1}\cdot\left(1+\frac{1}{n/\lambda-1}\right)$$

obsérvese que siendo λ una constante, cuando $n\to\infty$ también $(n/\lambda - 1)\to\infty$. Por tanto el primer factor de la última expresión es igual a *e* cuando $n\to\infty$. El segundo factor evidentemente es igual a 1 cuando $n\to\infty$.

Por tanto,

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n=\lim_{n\to\infty}\left[\left(1-\frac{1}{n/\lambda}\right)^{\frac{-n}{\lambda}}\right]^{-\lambda}=e^{-\lambda}$$

4. Respecto a la última parte resulta claro que $\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x}=1$

Recopilando, tenemos que en las condiciones indicadas de $n\to\infty$, $p\to 0$ con *np* constante e igual a λ, se verifica que:

$$P(X=x)=e^{-\lambda}\frac{\lambda^x}{x!}$$

18 ¿Cómo se puede ver que la distribución de la varianza muestral (S^2) está relacionada con la distribución chi-cuadrado?

Utilizaremos la definición de χ^2 con ν grados de libertad como $\sum_{i=1}^{\nu} z_i^2$ siendo las z_i variables aleatorias independientes con distribución $N(0; 1)$.

Respecto a la notación utilizada, consideraremos que $X \sim N(\mu;\sigma)$ y que $\bar{X}$, S^2 son la media y la varianza muestral de muestras aleatorias de tamaño n.

Podemos escribir que:

$$\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}\left[(X_i-\bar{X})+(\bar{X}-\mu)\right]^2 =$$

$$= \sum_{i=1}^{n}(X_i-\bar{X})^2 + 2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X}) + \sum_{i=1}^{n}(\bar{X}-\mu)^2$$

y como $\sum_{i=1}^{n}(X_i-\bar{X}) = 0$ y $(\bar{X}-\mu)^2$ es una constante, tenemos que:

$$\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2$$

Como $S^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2/(n-1)$ podemos escribir: $\sum_{i=1}^{n}(X_i-\bar{X})^2 = S^2(n-1)$.
Sustituyendo en la expresión anterior y dividiéndolo todo por σ^2 queda:

$$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} = \frac{S^2(n-1)}{\sigma^2} + \frac{(\bar{X}-\mu)^2}{\sigma^2/n}$$

Aceptando que los 2 sumandos que aparecen a la derecha en la expresión anterior corresponden a variables aleatorias independientes, y escribiéndolos de la forma

$$\sum_{i=1}^{n} z_i^2 = \frac{S^2(n-1)}{\sigma^2} + z^2$$

Dado que la suma de variables aleatorias independientes distribuidas según χ^2 sigue también esta misma distribución con un número de grados de libertad igual a la suma de los grados de libertad de las distribuciones sumadas, se espera que si $\sum_{i=1}^{n} z_i^2 \sim \chi_n^2$ y $z^2 \sim \chi_1^2$, entonces: $\frac{S^2(n-1)}{\sigma^2} \sim \chi_{n-1}^2$

19 ¿Por qué da un resultado distinto sumar *k* variables aleatorias de la misma distribución de probabilidad que tomar una y multiplicarla por *k*?

Efectivamente da un resultado distinto, aunque pueda parecer que esto va en contra de aquello que nos enseñaron en el colegio de que es lo mismo sumar una cantidad k veces que multiplicarla por k.

Da el mismo resultado cuando las cantidades a que nos referimos son magnitudes constantes. Para saber cuántos huevos hay en 8 docenas, como el número de huevos en una docena es una constante, podemos sumar 12 huevos 8 veces o multiplicar 12 por 8 (¡evidente!). Pero si lo que tenemos es una variable aleatoria, como el peso de un huevo, esto ya no es verdad. No es lo mismo el peso de una docena de huevos que el peso de un huevo multiplicado por 12.

No es lo mismo porque aunque las 2 variables obtenidas tienen el mismo valor medio, no tienen la misma variabilidad, y por tanto ya no podremos decir que sean iguales.

Para entender por qué esto es así, aclararemos en primer lugar el significado de la notación que vamos a utilizar. Llamaremos X a la variable aleatoria que consideramos (en nuestro ejemplo el peso de un huevo), podríamos decir que X se distribuye según una Normal, pero no necesitamos hacer referencia a ninguna distribución en concreto, sólo hay que tener claro que no es un valor fijo, sino una variable aleatoria. Si tomamos una docena, designaremos los pesos como X_1, X_2, ..., X_{12}. Cada una de las X_i es una variable aleatoria con idéntica distribución que X. En realidad son extracciones de la misma población, el subíndice solo indica el orden en que extraen.

Echando mano de las fórmulas correspondientes, y suponiendo que los 12 pesos son independientes, tendremos que la esperanza matemática y la varianza del peso de una docena será:

$$E(X_1 + X_2 + ... + X_{12}) = E(X_1) + E(X_2) + ... + E(X_{12}) = 12 \cdot E(X)$$

$$V(X_1 + X_2 + ... + X_{12}) = V(X_1) + V(X_2) + ... + V(X_{12}) = 12 \cdot V(X)$$

Pensemos ahora en la variable "peso de un huevo multiplicado por 12", es decir, la variable $12X$. Ahora tendremos:

$$E(12X) = 12 \cdot E(X)$$

$$V(12X) = 12^2 \cdot V(X) = 144 \cdot V(X)$$

Vamos a ver que estas fórmulas reflejan lo que ocurre en la realidad a través una mirada intuitiva al problema. Cuando formamos una docena de huevos, aunque tomemos uno singularmente pequeño no hay que temer que esta docena salga con un peso muy por debajo del valor medio, ya que seguramente también habrá otros con un peso por encima de la media, de forma que los más grandes compensarán el peso de los más pequeños y viceversa.

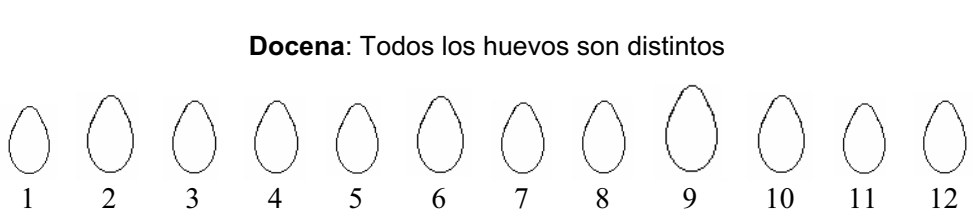


Figura 19.1. Peso de una docena de huevos. *En la docena habrá huevos más grandes de lo normal y otros más pequeños (en el dibujo se han exagerado las diferencias), y sus pesos tenderán a compensarse*

Sin embargo, en la variable "peso de un huevo multiplicado por 12", si el huevo elegido resulta ser pequeño, es como tener una docena de huevos pequeños; y si es grande sería como una docena de huevos grandes, con lo que tendremos pesos totales con más dispersión que en el caso anterior (porque en este caso no se compensan los grandes con los pequeños).

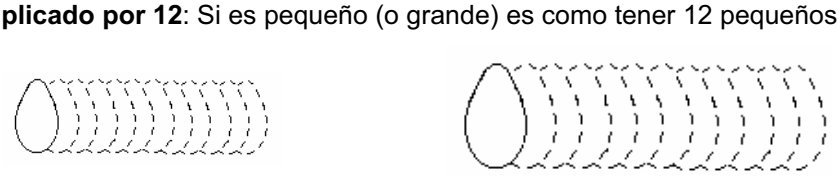


Figura 19.2. Peso de un huevo multiplicado por 12. *Si el huevo es singularmente pequeño es como tener una docena con todos los huevos pequeños. Igual si es grande*

Luego, tiene más dispersión la variable aleatoria "peso de un huevo multiplicado por 12" que la variable "peso de una docena". Y si la variabilidad es distinta, evidentemente las variables son distintas.

Esto nos obligará a estar atentos para no confundir lo que es la suma de k variables con el producto de una multiplicado por k (¡estamos tan acostumbrados a pensar que es lo mismo!). Para comprobar que sí sabe distinguir estas situaciones le sugerimos que piense en estas 4 que le planteamos.

a) El número de personas que suben por día a una atracción de feria es una variable aleatoria X. Si la atracción cuesta k, ¿cuánto vale la varianza de la recaudación diaria?

b) Se fabrican bolas de plástico cuyo volumen es una variable aleatoria Y, y su densidad es k. ¿Cuál es la varianza del peso de las bolas?

c) El valor de una resistencia eléctrica es una variable aleatoria Z. ¿Cuánto vale la varianza de la resistencia que resulta de conectar en serie k de estas unidades?

d) Unas barras metálicas tienen una longitud que es una variable aleatoria W. ¿Cuánto vale la varianza de la longitud de k de estas barras colocadas una a continuación de otra?[1]

[1] Soluciones: a) $k^2 \cdot \mathrm{Var}(X)$; b) $k^2 \cdot \mathrm{Var}(Y)$; c) $k \cdot \mathrm{Var}(Z)$; d) $k \cdot \mathrm{Var}(W)$

Estimación

20 Sabemos que las características de una muestra (proporción, media, ...) varían de una muestra a otra. ¿Por qué entonces creer en los resultados de una muestra, sabiendo que si tomáramos otra esos resultados serían distintos?

La respuesta a esta pregunta constituye la esencia de la teoría del muestreo estadístico y para responderla de una forma más concreta nos centraremos en la media. La clave está en que la media de la muestra, como muchos de los estimadores utilizados, tiene 2 importantes propiedades:

El valor medio de la media muestral coincide con la media de la población

La media varía de una muestra a otra, pero esta variabilidad se produce en torno al valor de la media poblacional. Es decir, si repetimos el proceso de muestreo, aunque los valores de las medias serán distintos, todos ellos se encontrarán agrupados (con mayor o menor dispersión) en torno a la media de la población. Cuando un estimador tiene esta propiedad se dice que es insesgado.

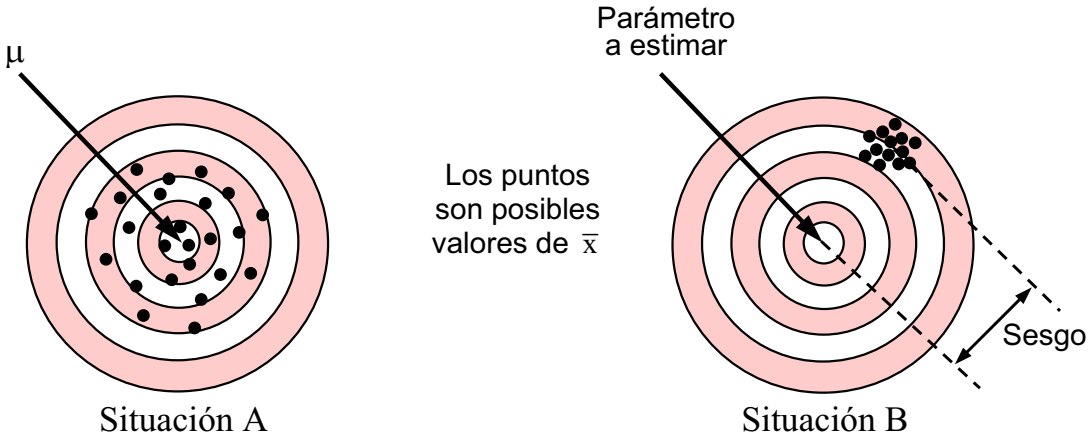


Figura 20.1. *Una mirada gráfica del insesgamiento de la media*

También es cierto que al tomar una sola muestra, conocemos un solo valor de la media muestral, y no tenemos manera de saber si se encuentra cerca o lejos de la media poblacional, pero para tratar con este inconveniente utilizamos la siguiente propiedad.

La variabilidad de la media disminuye al aumentar el tamaño de la muestra

Supongamos que la variabilidad de los datos viene definida por $\sigma=4$. Podemos calcular el número de elementos que debemos incluir en la muestra para que, por ejemplo, el 99,7% de las medias muestrales caigan a una distancia de menos de 2 unidades de la media poblacional (podría ser el círculo más pequeño de la diana).

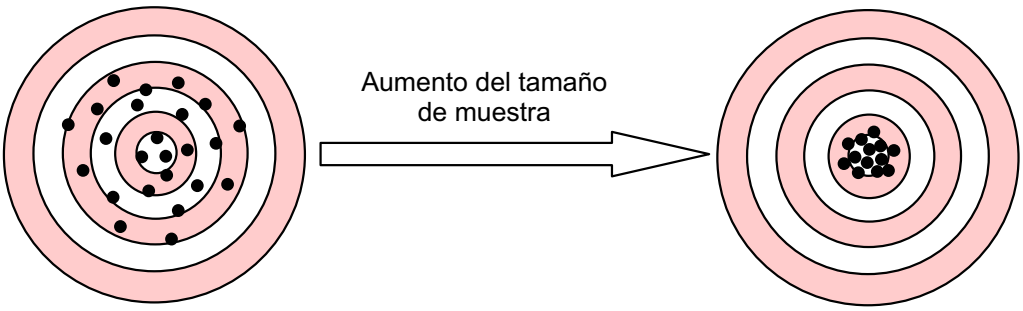


Figura 20.2. *Efecto del aumento del tamaño de la muestra en la reducción de la dispersión*

El tamaño de muestra requerido para cumplir con estas especificaciones es[1] n=36. Si se toman muchas muestras de este tamaño, ¿qué tan distintos serán los valores que presentarán sus medias? En este caso, el 99,7% de las veces caerán a una distancia de menos de 2 unidades de la verdadera media poblacional.

Está claro que en la práctica tomaremos una sola muestra, pero sabiendo que "casi" en todos los casos su media se encontrará a menos de 2 unidades de la media poblacional, esperamos que la obtenida por nosotros sea una de esas, aunque a decir verdad, no podremos asegurarlo, puesto que 3 de cada mil se salen de esa distancia. No tenemos más remedio que asumir un cierto riesgo de equivocarnos, aunque nosotros hemos podido fijar previamente la magnitud de ese riesgo. En nuestro ejemplo, equivale a comprar 997 boletos de una rifa que tiene 1.000 y que no nos toque el premio. Hay que tener muy mala suerte para que esto pase.

En resumen, siempre que en base a una muestra queramos conocer un rasgo de la población con un margen de error especificado, corremos un riesgo de equivocarnos, pero este riesgo se puede controlar, llevándolo al valor que convenga, utilizando el tamaño de muestra adecuado. Otra cosa es que dispongamos del presupuesto y de los medios necesarios para tomar una muestra aleatoria del tamaño necesario.

[1] La fórmula que permite calcular el tamaño de la muestra, cuando el tamaño de la población se supone infinito, o muy grande respecto al tamaño de la muestra, es: $n = z_{\alpha/2}^2 \sigma^2 / \delta^2$, donde el valor $z_{\alpha/2}$ depende del nivel de confianza (para el 99,7% vale 2,97), σ es la desviación tipo de la población y δ el margen de error. Si suponemos σ=4, el nivel de confianza indicado y δ=2, se obtiene n=35,2. Este valor conviene redondearlo por exceso.

21 ¿Qué significa la expresión: "un intervalo de confianza del 95% es 27,5% ± 3,6%"?

Esta expresión nos da un intervalo (27,5% ± 3,6%) que denominamos intervalo de confianza (en este caso es del 95%) y corresponde a la estimación de un porcentaje en cierta población. Podría ser, por ejemplo, el porcentaje de estudiantes de una universidad que usan el correo electrónico de forma habitual, según un estudio realizado a partir de una muestra. La verdad es que tantos porcentajes juntos aturden un poco, así que los comentaremos uno a uno, aunque no en el orden en que aparecen.

El 27,5% es el porcentaje que se tiene en la muestra y a este valor le llamamos estimación puntual. Por ejemplo, supongamos que la muestra está formada por 720 estudiantes y 198 responden afirmativamente, la proporción es 198/720 = 0,275, o lo que es lo mismo, 27,5%. Puestos a decir un número lo más razonable sería decir que la proporción en la población es del 27,5% o, como sabemos que si tomáramos otra muestra es casi seguro que la proporción sería distinta, es mejor decir que la proporción está "en torno al" 27,5%. Para cuantificar lo que significa ese "en torno al" se añaden los siguientes valores.

El ± 3,6% es lo que llamamos margen de error. Añadiendo este valor queremos decir que el porcentaje que estimamos no es "exactamente" el 27,5, sino un valor probablemente comprendido entre 23,9 y 31,1 (27,5% ± 3,6%).

Finalmente, el 95% es el nivel de confianza de la estimación. Significa que el intervalo que damos se ha construido mediante un procedimiento que garantiza que el 95% de los intervalos que se obtendrían, si repitiéramos el muestreo, atraparían el verdadero valor que andamos buscando. Este concepto requiere una cierta reflexión. No sabemos, no tenemos manera de saber, si en nuestro caso concreto el valor buscado está dentro o fuera del intervalo (si supiéramos que está dentro podríamos decir que la confianza es del 100%). En realidad, lo único que sabemos es que el intervalo se ha construido usando una metodología, aplicando unas fórmulas, que aciertan en el 95% de los casos. Es como si esta información nos la diera alguien que dice la verdad solo el 95% de las veces, seguramente la tomaríamos en consideración (especialmente si no tenemos nada más a que agarrarnos) aunque no podríamos estar seguros de que fuera cierta.

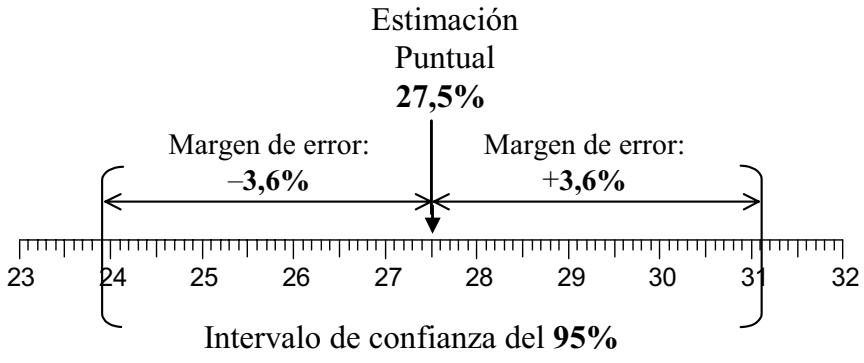


Figura 21.1. *Esquema gráfico de un intervalo de confianza del 95% para la estimación de un porcentaje*

Si una confianza del 95% no nos parece suficiente, podemos calcular el intervalo con una confianza del 99%. Dicho esto, si es posible tener un mayor nivel de confianza, ¿por qué no lo hacemos así ya desde el principio? Porque cuanta más confianza queramos más ancho resultará el intervalo (para estar más seguros de que el valor buscado está dentro, lo hacemos más ancho), y lo que ganamos en seguridad lo perdemos en precisión. Exagerando mucho, incluso sin tener datos, podríamos afirmar con una confianza del 100% que el porcentaje buscado está entre el 0 y el 100%, pero es evidente que esta afirmación –aun con tanta confianza– no sirve para nada.

Suponiendo que en nuestro ejemplo el número de estudiantes de la universidad (tamaño de la población) es N = 25.000 y con los otros datos que ya hemos ido comentando (tamaño de muestra n = 720, proporción en la muestra p = 0,275) se obtienen los siguientes intervalos para distintos valores de la confianza:

Nivel de confianza	Intervalo
90 %	23,5 – 30,5
95 %	23,9 – 31,1
99 %	22,8 – 32,2
99,9 %	21,5 – 33,5

Estos intervalos se calculan, considerando el cumplimiento de ciertas condiciones, mediante la fórmula[1]:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{N-n}{N}\cdot\frac{p(1-p)}{n-1}}$$

Donde $\hat{p}$ es la proporción calculada en la muestra, $z_{/2}$ un coeficiente relacionado con el nivel de confianza de la estimación (a mayor nivel de confianza, mayor valor de este coeficiente), N es el tamaño de la población, n el tamaño de la muestra y p la proporción que se anda buscando en la población. Como este último valor no se conoce, si se quiere una estimación conservadora se coloca el más desfavorable, el que da mayor margen de error, y es fácil comprobar que este es p = 0,5. De la fórmula también se puede deducir que una forma de disminuir el margen de error es aumentar el tamaño de la muestra n. Por tanto, es posible aumentar el nivel de confianza manteniendo el margen de error, a base de aumentar el tamaño de la muestra.

Es necesario tener en cuenta que este tipo de cálculos se basan en que la muestra es aleatoria, y en que las preguntas se han realizado de la forma adecuada para obtener la información que se desea. No debe pasar inadvertido que la pregunta que hemos utilizado en este ejemplo es un poco vaga, ya que no todo el mundo entiende lo mismo por usar el correo electrónico de forma habitual. Las imprecisiones relacionadas con la falta de aleatoriedad de la muestra, o con una forma poco concreta o descuidada de hacer las preguntas, suelen acarrear errores mucho mayores que aquel que figura en el intervalo de confianza, y que habitualmente es el único a que se hace referencia.

[1] Las condiciones son que $np > 5$ y $n(1-p) > 5$. La deducción de la fórmula puede encontrarse, por ejemplo, en el texto de M. García Ferrando: *Socioestadística. Introducción a la Estadística en Sociología*. Alianza Editorial, 1988. En muchos textos esta fórmula aparece sin el término $(N-n)/N$, ya que si el tamaño de la población es mucho mayor que el de la muestra (lo más frecuente) este cociente apenas afecta al resultado.

22 ¿Por qué para estimar la media de una población el tamaño de la muestra no crece proporcionalmente con el tamaño de la población?

Sobre el tamaño de una muestra se dicen muchas cosas en la cultura popular, no todas ellas ciertas. Por ejemplo, en ocasiones se rechazan los resultados de un estudio, con el argumento de que "la muestra no es representativa, pues ni siquiera llega a ser el 10% de la población". Esta cifra del 10%, o cualquier otra, es completamente falaz, y es conveniente desmitificarla a través de algunas situaciones como las que se comentan a continuación.

¿Le hace falta sal a la sopa?

Para responder esta pregunta el cocinero debe tomar una "muestra representativa" de sopa, probarla (medir la sal con sus papilas) y luego tomar una decisión para toda la sopa. ¿Le parece a usted sensato recomendar al chef que pruebe el 10% de la sopa para que pueda tomar una "buena decisión"? Esta recomendación contradice todas nuestras experiencias, pues la cucharilla que usamos en nuestras casas para probar la sopa que preparamos habitualmente es la misma que cuando la preparamos en una olla más grande porque tenemos invitados, y esto nunca nos ha sorprendido.

Figura 22.1. *El tamaño de la cucharilla no depende del tamaño de la olla*

Una acción que no se perdona antes de meter la cucharilla en la sopa es la de hacer un buen meneo con el cucharón para homogeneizar la sopa, de tal manera que cualquier posible muestra que se tome brinde la misma información que otra. Es decir, que en esta situación, como en casi todas, se considera que es más importante la variabilidad o la homogeneidad, que el tamaño de la olla. Dicho de otra manera, es mucho más importante mezclar bien antes de tomar la muestra, que aumentar el tamaño de la cuchara.

Doctor, ¿para saber mi grupo sanguíneo me va usted a sacar una muestra con el 10% de mi sangre?

Es bien sabido que basta con una muestra de una gota de sangre para conocer en forma inequívoca el grupo sanguíneo de una persona; esto, en virtud de que todas las gotas de sangre de una persona son del mismo tipo, así que vista una, quedan vistas todas. De nuevo vemos aquí que el impacto de la homogeneidad es más importante que el del tamaño de la población. La misma cantidad de sangre se requiere para un niño recién nacido que para su padre.

¿Para saber en promedio cuántas patas tienen los cangrejos del Pacífico, necesito capturar el 10%?

De nuevo aquí se desmitifica la falacia de que una muestra, para que sea "representativa", debe contener una cantidad de unidades de acuerdo con el tamaño de la población hacia la que se desea concluir.

En este caso hay un agravante. ¿Quién nos dirá cuánto es el 10% de los cangrejos del Pacífico?

Bueno, ya que hemos recurrido a la intuición, ahora vamos a mostrar el verdadero efecto del tamaño "N" de una población en la definición del tamaño "*n*" que debe tener una muestra para realizar estimaciones con una calidad previamente especificada.

Relación del tamaño *N* de una población con el tamaño *n* que deberá tener una muestra

La pregunta ¿cuál debe ser el tamaño de muestra a tomar? se puede formular técnicamente de la siguiente forma: si queremos estimar la media μ de una población que tiene *N* unidades, con un margen de error δ y con un nivel de confianza $(1-\alpha)$, ¿cuál debe ser el tamaño *n* de muestra que debemos tomar?

Después de hacer algunas suposiciones y algo de álgebra para despejar el valor de *n*, se obtiene la siguiente fórmula[1] en la que $z_{\alpha/2}$ es función del nivel de confianza, y σ es la desviación tipo de la población.

$$n = \frac{\dfrac{z_{\alpha/2}^2 \sigma^2}{\delta^2}}{1 + \dfrac{1}{N}\left(\dfrac{z_{\alpha/2}^2 \sigma^2}{\delta^2}\right)}$$

[1] La expresión del intervalo de confianza $1-\alpha$ para μ, con la corrección para poblaciones finitas es $\bar{x} \pm z_{\alpha/2} \cdot \dfrac{\sigma}{\sqrt{n}} \sqrt{\dfrac{N-n}{N}}$. Por tanto, $\delta = z_{\alpha/2} \cdot \dfrac{\sigma}{\sqrt{n}} \sqrt{\dfrac{N-n}{N}}$. A partir de aquí se puede despejar *n*.

Para la estimación definida por δ, (1-α), y σ, la cantidad del numerador es constante (no depende de N), llamémosle n_0. De esta manera la expresión anterior se convierte en:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Observe que cuando el tamaño de la población N se hace muy grande, la proporción n_0/N tiende a cero y por lo tanto $n = n_0$ (en términos matemáticos se diría que cuando N tiende a infinito). En este caso, el tamaño de la muestra puede calcularse como: $n_0 = z^2_{\alpha/2}\sigma^2 / \delta^2$ que no depende del tamaño de la población.

Supongamos que estamos dispuestos a tolerar un margen de error δ = 1, la desviación tipo de la población es σ = 8 y el nivel de confianza deseado es del 95%, con lo cual, en este caso, tendremos $z_{0,025}$=1,96. Con estos datos, n_0 = 246, veamos cuál es el tamaño de la muestra si el tamaño de la población es N = 500.

$$\text{Si } N = 500 \Rightarrow n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{246}{1 + \frac{246}{500}} = \frac{246}{1 + 0,492} = \frac{246}{1,492} = 164,88 \simeq 165$$

En la Tabla 22.1 se repite este cálculo para distintos tamaños de N y se aprecia muy claramente cómo a medida que crece la población, su tamaño pierde influencia sobre el tamaño de la muestra. Vemos cómo cuando la población cambia de tamaño N=100 a N=150, solo 50 unidades, el cambio de la muestra va de n=72 a n=94, es decir, que se aumenta en 22 unidades, sin embargo, cuando el tamaño de la población cambia de N=5.000 a N=10.000, es decir, aumenta en 5.000 unidades, el tamaño de la muestra cambia en tan solo 5 unidades.

Tabla 22.1. *Impacto del tamaño de la población sobre el tamaño de la muestra para estimar la media con una confianza del 95%, un margen de error δ =1 y una desviación tipo de la población σ =8*

Tamaño N de la población	Tamaño n de la muestra
100	72
150	94
200	111
250	124
500	165
1.000	198
2.000	219
5.000	235
10.000	240
50.000	245
100.000	246
1.000.000	246
50.000.000	246

Cuando la población cambia 100.000 a 50.000.000, el tamaño de muestra, después de hacer el redondeo, se mantiene constante e igual a 246 unidades.

La Figura 22.2 muestra gráficamente la relación entre tamaño de la población y tamaño de la muestra.

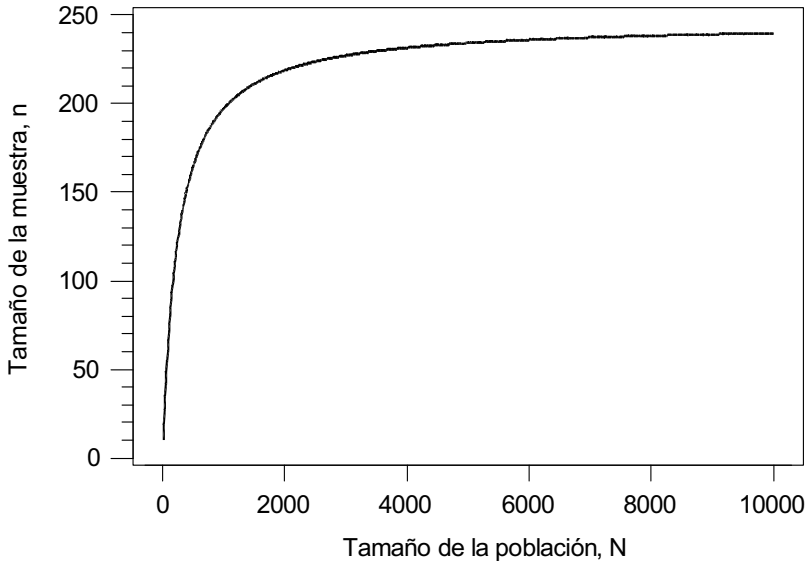


Figura 22.2. *Relación entre el tamaño de la población y el tamaño de la muestra para una confianza del 95%, un margen de error $\delta = 1$ y desviación tipo de la población $\sigma = 8$*

Nótese la pendiente tan pronunciada en el primer tramo, lo cual indica que cuando el tamaño de la población no es muy grande, este tiene gran impacto sobre el tamaño de la muestra, pero cuando el tamaño de la población es grande, ya no es tan importante, al punto de que el tamaño más grande posible para la muestra en la figura es n=246, para cualquier tamaño de la población.

23 ¿Por qué cuesta acertar en los sondeos electorales?

Los sondeos electorales son una de las aplicaciones estadísticas de la que más se habla (y es verdad que no siempre bien). Además, este tipo de estudios resulta singular desde varios puntos de vista, por ejemplo:

- Es un tema que despierta gran interés, incluso pasión, en gran parte de la población.
- A diferencia de otros estudios estadísticos, en este caso se acaba sabiendo el verdadero valor de los parámetros estimados, cosa que no ocurre si, por ejemplo, hacemos una encuesta para conocer el porcentaje de estudiantes que tienen conexión a internet en su casa
- Los sondeos electorales fallan con frecuencia (aunque no siempre), y como esta es la única relación de muchas personas con la estadística, se abona la impresión de que esta es una ciencia poco seria.

Siempre que se estiman las características de una población a partir de una muestra se corre un cierto riesgo de error, pero sabemos que este riesgo es controlable mediante la selección de un adecuado tamaño de la muestra. Ahora bien, cuando los errores son de bulto y se producen de forma repetitiva, es que fallan otras cosas. No es que no se cumpla la teoría estadística, lo que suele ocurrir es que no se han aplicado bien los principios en que se basa, ya sea por falta de recursos o porque dadas las circunstancias en que se realizan, es muy difícil cumplirlos. Veamos algunos casos.

La muestra no es representativa

Para la correcta predicción de lo que pasa en la población a partir de una muestra es fundamental que la muestra sea representativa (esto es evidente, ya que de lo contrario las conclusiones no se pueden extrapolar). La representatividad de la muestra se consigue mediante una adecuada selección de sus componentes, de forma que todos los individuos de la población tengan una probabilidad conocida de ser incluidos (la misma probabilidad en el caso de muestras aleatorias simples). Pero esto no es fácil. Ni barato.

Tal vez el fiasco más famoso en un sondeo electoral se produjo en las elecciones presidenciales de los Estados Unidos en 1936, cuando se enfrentaban los candidatos Roosevelt (demócrata) y Landon (republicano). La revista *Literary Digest* realizó un sondeo mediante el envío por correo de 10 millones de cuestionarios, de los cuales recibió cumplimentados 2,4 millones. Pero a pesar de lo generoso del tamaño muestral, la revista predijo una clara victoria de Landon (con el 57% de los votos) cuando en realidad ocurrió todo lo contrario (ganó Roosevelt con el 61% frente al 37% de Landon). El problema estaba en la selección de la muestra, ya que las direcciones a las que se enviaron los cuestionarios se obtuvieron de la guía telefónica y del registro de propietarios de automóviles. Este fue un gran error, pues en aquella época tener teléfono o coche estaba muy relacionado con la clase socioeconómica a la que se pertenecía, de forma que los partidarios de Landon estaban sobredimensionados en la muestra, en detrimento de los de Roosevelt.

Hoy en día, seleccionar a los integrantes de la muestra a través de la guía telefónica es mucho menos problemático porque el uso del teléfono se ha extendido a todas las capas de la población en los países desarrollados tecnológicamente. Pero en el caso de realizar la entrevista por teléfono (caso frecuente) sí tiene mucha importancia a qué hora se llama, por quién se pregunta, o cómo se sustituye a los que no desean contestar. Descuidar estos aspectos puede conducir a graves errores en las predicciones por sesgo en la selección de la muestra.

La intención de voto va cambiando

Los sondeos electorales se basan en encuestas realizadas varios días, o incluso varias semanas, antes de las elecciones. En algunos países está prohibido publicar resultados de sondeos electorales durante un cierto periodo de tiempo antes de las elecciones (en España este periodo es de una semana). Por tanto, nos encontramos con 2 tipos de extrapolaciones:

- La que se hace desde la muestra hacia la población, siendo esta la que trata la teoría estadística del muestreo.
- La que generaliza los resultados de las fechas en que se ha hecho el sondeo, al día de las elecciones. Pero los partidos se emplean a fondo en sus campañas electorales (quizá algunas mejor pensadas que otras), se producen debates entre candidatos, pueden ocurrir sucesos sobre los que se posicionan los candidatos, y todo esto puede afectar al voto decidido o a la decisión final de los que en el momento de la encuesta estaban indecisos.

Un caso paradigmático de cambio en la intención de voto (o quizá mejor, de cambio en la intención de abstenerse decidiendo ir a votar) es el que se produjo en las elecciones españolas de 2004. La getión informativa que hizo el gobierno del Partido Popular sobre el atentado del 11 de marzo (solo 3 días antes de las elecciones) indignó y movilizó a muchos electores de forma que, en contra de los pronósticos y sondeos que se venían realizando, ganó el Partido Socialista con una clara mayoría. En definitiva, los cambios en la intención de voto, si se producen, no se pueden predecir estadísticamente con base en las encuestas realizadas.

¿A quién votarán los indecisos?

Los indecisos son un verdadero dolor de cabeza para los encargados de realizar sondeos electorales, especialmente en los casos en que este grupo representa un porcentaje grande y cuando no hay un candidato que lleve una clara ventaja sobre el resto.

El problema está en que las personas que no responden no son una muestra al azar de la población. Si las razones de no respuesta fueran estadísticamente independientes de las preferencias de voto, el problema tendría solución desde la estadística. Sin embargo, la práctica ha demostrado que esto no es así, y el problema se complica

Fermín Bouza, en un interesante artículo publicado en la revista 'Praxis Sociológica'[1], escribe: *"…. Como quiera que el grupo de NS/NC [No Sabe/ No Contesta] suele ser*

[1] Fermín Bouza Álvarez: "Comunicación política: encuestas, agendas y procesos cognitivos electorales". Revista '*Praxis Sociológica*' número 3 (1998).

muy amplio en vísperas electorales no muy inmediatas (y, a veces, también en las inmediatas), y suele ir de un 20% a un 50% de los encuestados, el sociólogo, presionado por medios y partidos, tiene que hacer una estimación del voto, es decir, imputar a los indecisos un voto y predecir lo que ocurriría hoy si se celebraran las elecciones (o el día de las elecciones, lo cual es mucho más circense, todavía). Para hacer esa imputación a los indecisos cuenta con varios procedimientos, siendo los más frecuentes la atribución por simpatía ("¿Con qué partido simpatiza usted más?" o "¿De qué partido se siente usted más cercano?", o cualquier otra fórmula) o por recuerdo de voto (¨¿A qué partido votó usted en las últimas elecciones?"). No son las únicas fórmulas: los sociólogos pueden improvisar otras preguntas para conseguir la intención más probable de voto, y aquí ponen en juego su conocimiento, su intuición, o cualquier otra virtud adivinatoria. […] Lo que el sociólogo tiene delante es un 'menú' de estimaciones entre las que va a elegir una para dar a la opinión pública".

Está claro que la asignación de los votos de los indecisos a uno u otro partido es una tarea de importancia crítica, y su éxito responde más a conocimientos relacionados con la Sociología y con la Política que con la Estadística.

La falta de sinceridad en las respuestas

La redacción de las preguntas y el orden en que se realizan son también aspectos críticos. Escribir preguntas claras, que no induzcan respuestas, no es tarea fácil y requiere conocer bien la técnica de cómo plantear las preguntas y también requiere entrevistadores bien entrenados y motivados (léase bien pagados).

Puede preguntarse dando el nombre del candidato y el partido al que pertenece o dando solamente el nombre. Y esta pregunta puede hacerse antes o después de preguntas relacionadas con la situación del país, o con la valoración que se da a los candidatos, y las respuestas pueden variar dependiendo de cómo se haga la pregunta.

A veces existen condiciones de libertad de expresión particulares, que hacen más o menos creíbles las respuestas de los ciudadanos y que harán que el volumen de los llamados "indecisos" sea mayor o menor, ya que quizá en vez de indecisos lo que se tiene son "decisos cautos", que prefieren mantener reservada su opinión.

Del porcentaje de votos al número de escaños

En muchas ocasiones lo verdaderamente relevante, más que el porcentaje de votos que va a obtener cada partido, es el número de escaños, y los sistemas que se utilizan para distribuir los escaños en función del porcentaje de votos (como la ley d'Hondt), acaban de complicar las cosas. Por ejemplo, para una determinada circunscripción electoral en la que hay 5 escaños en juego se puede predecir con una confianza del 95% que un determinado partido obtendrá un 32% de los votos con un margen de error del 3%. El problema está en que si obtiene un 31% le corresponderá un escaño, mientras que si obtiene el 33% le corresponderán 2. Y esta es una diferencia importante pero no sabemos por cuál opción decantarnos con la información disponible.

Otro problema es que algunas legislaciones electorales exigen un mínimo porcentaje de votos (por ejemplo, el 5%) para entrar en el reparto. Si un partido está rozando este

porcentaje (por ejemplo, si su estimación de voto es del 4±2% no se puede saber si llegará o no, y el hecho de que ocurra una cosa u otra afectará también al número de escaños del resto de partidos.

A modo de resumen

Cuando se realizan sondeos electorales existen muchas dificultades para lograr buenas predicciones, dificultades que van más allá de aquellas que se refieren al ámbito de la teoría del muestreo estadístico (por no hablar de manipulaciones y de resultados interesados). A pesar de ello también se producen notables éxitos en las predicciones, o en el adelanto de los resultados muy poco tiempo después de cerrarse las urnas. Un interesante ejemplo a este respecto puede verse en un artículo de J.M. Bernardo[2] donde describe las técnicas utilizadas para seguir las intenciones de voto y también para poder anunciar solo 2 horas después de cerrarse las urnas, horas antes de cualquier otra aproximación provisional a los resultados, que el partido socialista había obtenido 201 de los 350 escaños del parlamento en las elecciones generales españolas de 1982 (el resultado oficial, conocido una semana después, fue que había obtenido 202).

Sería conveniente tener medida la frecuencia y la magnitud con que fallan los sondeos electorales serios (de los otros no hablamos), pues de la misma manera que las malas noticias son las que nos invaden a través de los medios informativos, también las pifias en las predicciones son las más destacadas, incluso en el ambiente académico, pues es más sensacional y a veces más pedagógico ilustrar lo que no debe hacerse, que mostrar ejemplos donde las predicciones han funcionado bien.

Debe quedar claro que en este caso de los sondeos electorales hay varios tipos de procesos involucrados, y en muchos de ellos la sociología, la psicología o la politología juegan un papel más protagonista que la propia estadística. En cualquier caso, el uso apropiado de las herramientas estadísticas, la seriedad con que se aborde el trabajo de campo, la supervisión, el conocimiento sociológico del grupo humano de interés, son factores claves en el éxito de las predicciones.

También pueden existir, y de hecho existen, encuestas que son el resultado de consultas interesadas que pretenden influir sobre la opinión de los electores. El ciudadano, tendrá que aprender a distinguirlas, aunque a veces es difícil por la proliferación y bombardeo de resultados de encuestas y sondeos. Indagar sobre el patrocinador y responsable del estudio puede dar cuenta del interés por la divulgación de determinados resultados, sin que ello obligue a sospechar de sesgo o falsedad. La experiencia y seriedad de la agencia responsable del estudio, así como del medio en que se publica, también son un buen indicador de la confianza que merecen los sondeos, además de aquella que se indica en la ficha técnica.

[2] José M. Bernardo "Monitoring the 1982 Spanish Socialist Victory: A Bayesian Analysis". Journal of the American Statistical Association. Vol. 79, Núm. 387 (1984).

24 ¿Qué es un estimador de máxima verosimilitud?

Cuando queremos estimar un parámetro de cierta población de la que sabemos la familia a que pertenece (Normal, binomial, Poisson, ...), tomamos una muestra aleatoria para a partir de ella construir un estadístico que permita estimar dicho parámetro. Por ejemplo, si se sabe que la situación particular podría modelarse con una distribución de Poisson, tenemos que:

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}; \quad x = 0,1,2....$$

Con la intención de aproximarnos a conocer el valor de λ, podríamos tomar una muestra aleatoria (x_1, x_2, ... x_n) de *n* observaciones e intentar responder a la siguiente pregunta ¿Cuál de todas las posibles poblaciones que define el parámetro λ, produce con mayor probabilidad una muestra como la observada?, o lo que es lo mismo, ¿cuál de todos los posibles valores de λ, produce con mayor verosimilitud (credibilidad) una muestra como la observada?. La respuesta a esta pregunta corresponde al estimador de máxima verosimilitud.

Para hacer operativo el criterio descrito vamos a deducir la expresión del estimador máximoverosimil de λ en nuestro ejemplo de la distribución de Poisson. La matemática es quizá un poco aparatosa, pero creemos que será fácil de seguir para las personas interesadas.

Consideremos la muestra aleatoria (x_1, x_2, ... x_n), que puede ser obtenida con una probabilidad: $P(X_1 = x_1; X_2 = x_2; ... X_n = x_n)$. Como las variables aleatorias que componen la muestra son independientes, la probabilidad conjunta es igual al producto de las probabilidades de cada una de ellas, de manera que:

$$L = P\left(X_1 = x_1, X_2 = x_2, ... X_n = x_n \middle| \lambda\right)$$

$$= \underbrace{P\left(X_1 = x_1 \middle| \lambda\right)}_{\frac{e^{-\lambda}\lambda^{x_1}}{x_1!}} \cdot \underbrace{P\left(X_2 = x_2 \middle| \lambda\right)}_{\frac{e^{-\lambda}\lambda^{x_2}}{x_2!}} \cdot ... \cdot \underbrace{P\left(X_n = x_n \middle| \lambda\right)}_{\frac{e^{-\lambda}\lambda^{x_n}}{x_n!}}$$

La expresión que hemos nombrado con la letra *L*, es una función de λ, pues se supone que (x_1, x_2, ... x_n) es una realización de la muestra, es decir, son números concretos. Así que en *L*, el parámetro λ es nuestra única incógnita. Esto podemos escribirlo así:

$$L\left(\lambda \middle| x_1, x_2, ... x_n\right) = \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdot \frac{e^{-\lambda}\lambda^{x_2}}{x_2!} \cdot ... \cdot \frac{e^{-\lambda}\lambda^{x_n}}{x_n!}$$

La pregunta ahora puede formularse de la siguiente manera: Si se obtuvo la muestra (x_1, x_2, ... x_n), ¿cuál es el valor de λ (en términos de las x_i) que hace máxima la función $L(\lambda \mid x_1, x_2, ... x_n)$?

De esta manera el problema queda convertido en un problema de optimización, del tipo de los que resolvimos en nuestros cursos de cálculo. De aquellos en los que debía derivarse, igualar a cero la derivada, para encontrar puntos críticos que luego serían probados con la segunda derivada para identificar si correspondían con un punto máximo o con un mínimo.

Para que la deducción sea más cómoda, un truco que suele dar buenos resultados es tratar con el logaritmo de L en vez de con L directamente, ya que el valor de λ que maximice una expresión, maximizará también la otra. En nuestro caso lo haremos de esta forma, obteniéndose:

$$\ln\left[L(\lambda)\right] = \ln\left(e^{-\lambda}\frac{\lambda^{x_1}}{x_1!}\cdot e^{-\lambda}\frac{\lambda^{x_2}}{x_2!}\cdot \ldots \cdot e^{-\lambda}\frac{\lambda^{x_n}}{x_n!}\right) =$$

$$= \ln\left(e^{-\lambda}\frac{\lambda^{x_1}}{x_1!}\right) + \ln\left(e^{-\lambda}\frac{\lambda^{x_2}}{x_2!}\right) + \ldots + \ln\left(e^{-\lambda}\frac{\lambda^{x_n}}{x_n!}\right) =$$

$$= \left[-\lambda + x_1\ln(\lambda) - \ln(x_1!)\right] + \left[-\lambda + x_2\ln(\lambda) - \ln(x_2!)\right] + \ldots$$

$$+ \left[-\lambda + x_n\ln(\lambda) - \ln(x_n!)\right] = -n\lambda + \ln(\lambda)\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\ln(x_i!)$$

Derivando respecto a λ e igualando a cero:

$$\frac{\partial \ln\left[L(\lambda)\right]}{\partial\lambda} = -n + \frac{1}{\lambda}\cdot\sum_{i=1}^{n} x_i = 0$$

para asegurar que estamos ante un máximo, hacemos:

$$\frac{\partial^2 \ln\left[L(\lambda)\right]}{\partial\lambda^2} = -\frac{1}{\lambda^2}\cdot\sum_{i=1}^{n} x_i < 0$$

De la primera derivada se deduce fácilmente que el valor de λ que maximiza ln[L(λ)], y por tanto también L(λ) es:

$$\lambda = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$$

Por tanto, diremos que la media de la muestra es el estimador de máxima verosimilitud para el parámetro λ de una distribución de Poisson.

Contraste de hipótesis

25 ¿Qué es el p-valor y cuál es el significado de las otras palabras clave que aparecen en el contraste de hipótesis?

Vamos a utilizar un ejemplo, que servirá para presentar el concepto de p-valor y para mostrar un panorama general de la forma de razonamiento que llamamos "contraste de hipótesis". Aunque en este tipo de ejemplos lo más habitual es tratar sobre la estimación de la media poblacional (quizá porque es conceptualmente sencillo y de mucha aplicación), aquí vamos a cambiar de protagonista planteando un caso que gira en torno al coeficiente de correlación[1].

La situación es la siguiente: un producto que recientemente se ha empezado a fabricar se obtiene mediante una reacción química. *A grosso modo*, lo que se hace es introducir en el reactor las materias primas, se calientan hasta una cierta temperatura y se deja que reaccionen. El problema está en que no se logra que reaccionen completamente y la parte que queda sin reaccionar, como se ha ensuciado y ha sufrido algunas transformaciones, no se puede volver a utilizar y hay que tirarla. El porcentaje de producto que se obtiene respecto al que se podría obtener es lo que llamamos rendimiento de la reacción, y en las 20 veces que se ha realizado hasta ahora, el valor obtenido está en torno al 75%.

Se sospecha que aumentar la temperatura tiende a aumentar el rendimiento, pero no se quiere cambiar sin estar razonablemente seguros de que esto es así (entre otras razones porque en caso de no tener efecto aumentarían innecesariamente los costes de energía). El valor que está estandarizado para la temperatura es de 110°C, pero por diversas razones (básicamente variabilidad en los aparatos de regulación) no se consigue que la reacción se produzca siempre a este valor.

Como cada vez que se ha llevado a cabo la reacción se ha anotado el rendimiento junto con los parámetros de producción más relevantes, entre ellos la temperatura, se ha podido construir el diagrama de la Figura 25.1

Algunas personas pueden pensar que la relación entre rendimiento y temperatura es evidente. Pero a otras esta relación les parecerá no tan clara. También se puede pensar que aunque el diagrama parece mostrar cierta relación, si se tomaran los datos de otras 20 reacciones su aspecto podría ser completamente distinto, por lo que no se puede considerar que los datos existentes demuestren nada de forma clara.

Para aclarar este dilema, vamos a empezar cuantificando la relación entre ambas variables a través de su coeficiente de correlación. En este caso tenemos que su valor es $r = 0{,}53$.

[1] No debe preocuparse el lector si no está familiarizado con el coeficiente de correlación, ya que podrá seguir las explicaciones perfectamente. Para su información, el coeficiente de correlación es una medida de la relación lineal que existe entre 2 variables y tiene la gran ventaja de que sus valores siempre están comprendidos entre –1 (puntos alineados en una recta de forma que al aumentar X disminuye Y) y 1 (los puntos también está alineados perfectamente según una recta, pero al aumentar X también aumenta Y). Valores del coeficiente de correlación "alrededor del cero" indican que no puede afirmarse que exista relación lineal entre las 2 variables. Lo que significa "alrededor del cero" es el quid de la cuestión en la pregunta que estamos respondiendo.

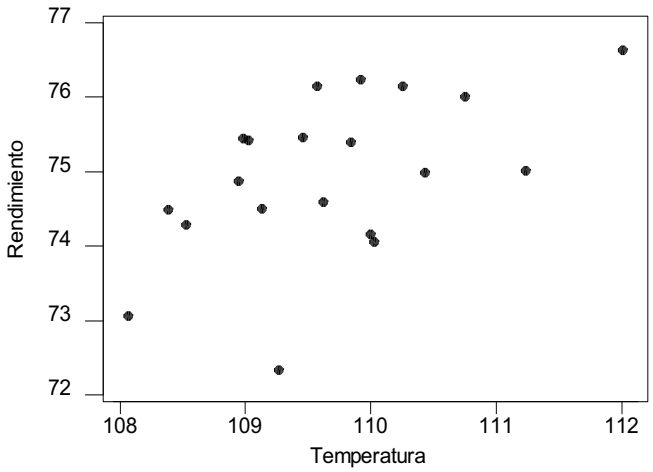


Figura 25.1. *Diagrama bivariante con los valores de temperatura y rendimiento obtenidos las 20 veces que se ha realizado la reacción*

El hecho de que este valor sea mayor que cero ¿nos permite afirmar que existe relación? Pues la verdad es que no, porque si 2 variables son independientes (no existe ninguna relación entre ellas), no por eso van a tener un coeficiente de correlación muestral exactamente igual a cero. Para ponerlo de manifiesto hemos simulado 20 parejas de números aleatorios (X, Y). Los valores de X se han obtenido de una distribución Normal con media 110 y desviación tipo 1, y los de Y también de una distribución Normal con media 75 y desviación tipo 1 (suponemos que los valores de la temperatura y el rendimiento se pueden considerar pertenecientes a estas distribuciones)[2]. Los dos conjuntos de datos no tienen nada que ver entre sí, y calculado el coeficiente de correlación entre X e Y hemos obtenido el valor de $r = -0{,}19$. El diagrama bivariante de estos valores es el representado en la Figura 25.2.

En el problema que nos ocupa la clave está en preguntarse: ¿es el valor obtenido para el coeficiente de correlación suficientemente distinto de cero como para poder afirmar que efectivamente la relación existe? o, dicho en otras palabras, ¿podemos decir que el coeficiente de correlación es significativamente distinto de cero?

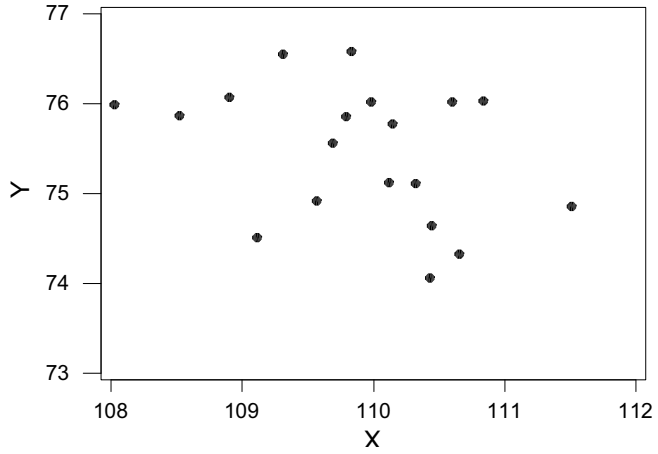


Figura 25.2. *Diagrama bivariante correspondiente a 20 pares de datos independientes*

[2] Aunque en realidad, para el uso que vamos a hacer de los resultados obtenidos en esta simulación, no es necesario que los valores generados pertenezcan a las distribuciones que hemos supuesto para rendimiento y temperatura.

Para responder a esta pregunta vamos a repetir 10.000 veces lo que hemos hecho para construir el diagrama de la Figura 25.2. Es decir, generaremos 20 números aleatorios simulando 20 valores de la temperatura y a continuación haremos lo mismo para 20 valores del rendimiento. Calcularemos su coeficiente de correlación y lo guardaremos. Al final tendremos 10.000 valores del coeficiente de correlación y cada uno de ellos corresponderá a 20 parejas de datos independientes. Los valores obtenidos en nuestra simulación los hemos representado en forma de histograma en la Figura 25.3

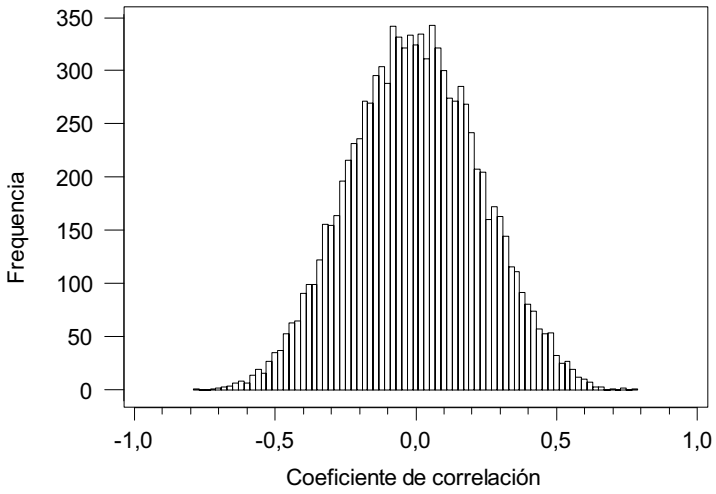


Figura 25.3. *Histograma de 10.000 valores del coeficiente de correlación obtenidos de muestras independientes de tamaño n = 20*

Ahora podemos razonar de la siguiente forma: Si no hubiera relación entre temperatura y rendimiento (coeficiente de correlación poblacional ⇐0) el valor del coeficiente de correlación muestral obtenido pertenecería a la distribución representada en la Figura 25.3. Si en nuestro caso el coeficiente fuera 0,2 esto no demostraría que existe relación entre ambas variables, ya que un valor como este, o incluso mayor, se da con mucha frecuencia cuando no hay relación. Si el valor fuera 0,85 tendríamos buenas razones para dudar de que el coeficiente de correlación poblacional fuera ⇐0, puesto que este valor es muy difícil que se produzca en ese caso (en nuestra simulación de 10.000 valores no ha salido ni una sola vez).

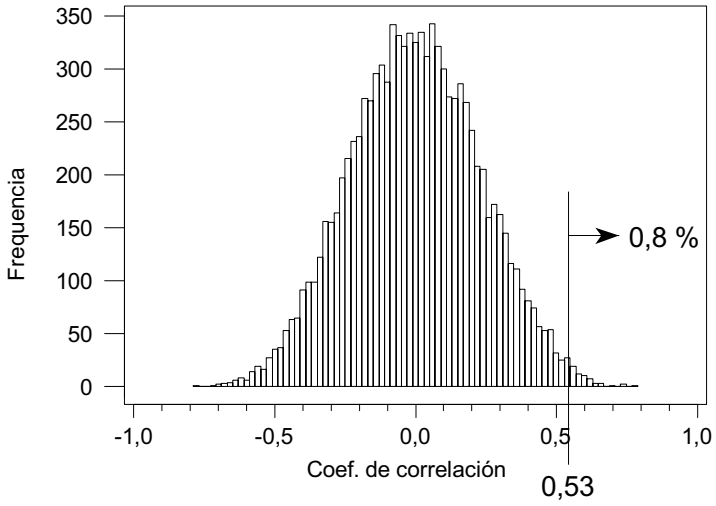


Figura 25.4. *Comparación del estadístico de prueba con la distribución de referencia*

En el ejemplo que estamos viendo, el coeficiente de correlación es 0,53 y lo que haremos es determinar la proporción de valores obtenidos que son iguales o mayores[3] que 0,53. Ordenando los 10.000 valores observamos que esta proporción es 0,0080.

¿Qué podemos decir? Pues que si no hubiera relación entre ambas variables, un valor del coeficiente de correlación tan grande como el que nos ha salido, o mayor, se daría del orden del 0,8% de las veces. ¿Es el valor 0,8% suficientemente pequeño para poder afirmar que hay relación entre las variables? La respuesta depende del riesgo de equivocarnos que estemos dispuestos a asumir considerando que hay relación cuando en realidad no la hay. Este riesgo se fija previamente y si, por ejemplo, estamos dispuestos a que sea del 5% como máximo, como nuestra probabilidad (0,8%) es menor, diremos que existe suficiente evidencia en los datos para dudar de la hipótesis de que no existe correlación entre las dos características.

Ahora vamos a recapitular y a reflexionar sobre nuestra forma de razonamiento. Hemos empezado diciendo que cambiaremos el valor de la temperatura si los datos de que disponemos ponen de manifiesto "de una forma razonable" que aumentar la temperatura aumenta el rendimiento. Formalmente este planteamiento se realiza bajo la forma de una *hipótesis nula,* que es la hipótesis más conservadora, la hipótesis de que el cambio que se pretende introducir no tiene efecto ("es nulo") frente a la *hipotesis alternativa*, que es lo que suponemos que ocurre si la hipótesis nula es falsa, en nuestro caso, que aumentar la temperatura aumenta el rendimiento.

A partir de los datos hemos calculado un valor que nos sirve para resumir en un solo número la información disponible. En nuestro caso este valor ha sido el coeficiente de correlación. En general a este valor se le denomina *estadístico de prueba.*

A continuación hemos deducido cuál es la distribución a que pertenecería el estadístico de prueba si la hipótesis nula fuera cierta. A esta distribución le llamamos distribución de referencia y en nuestro caso la hemos deducido por simulación[4].

Comparando el valor obtenido del estadístico de prueba con la distribución de referencia hemos calculado la probabilidad de que se dé un valor como el obtenido o mayor en el caso de que la hipótesis nula sea cierta. Al valor de esta probabilidad se le llama "valor p", "*p value*" o, tal como le hemos denominado nosotros, "*p-valor*".

El p-valor nos informa sobre el grado de compatibilidad de nuestros datos con la hipótesis nula. Así, cuando el p-valor es grande, por ejemplo 25%, entendemos que nuestra muestra no nos proporciona argumentos para dudar de la hipótesis nula. Cuando el p-valor es muy pequeño puede interpretarse como un indicador de incompatibilidad entre la hipótesis nula y los datos observados, pues estaría diciendo que si la hipótesis nula fuera cierta, sería muy raro obtener unos datos como los que obtuvimos.

En este contexto hay que definir el significado de "raro" en términos de probabilidad de ocurrencia. Así, por ejemplo, podríamos decir que raro es aquello que ocurre menos del 2% de las veces y este sería el nivel de referencia para comparar con el p-valor. A este valor de referencia le llamamos *nivel de significación* y nos referimos a él con la letra griega α.

---

[3] Iguales o mayores porque nuestro objetivo es analizar si existe correlación positiva. Si se tratara de ver si existe algún tipo de correlación (ya sea positiva o negativa), habrían que contar en cuantos casos se ha obtenido una diferencia respecto al cero, positiva o negativa, como la nuestra. En este caso la proporción sería el doble, ya que la distribución es simétrica.

[4] Algunas distribuciones de referencia son distribuciones teóricas muy bien conocidas, como la *t* de *Student*, la Chi cuadrado, etc. También para el coeficiente de correlación la distribución teórica es conocida cuando la distribución de la variable (X, Y) es Normal bivariante.

26 ¿A partir de qué p-valor es razonable rechazar la hipótesis nula?

Lamentablemente, aunque a todos nos gustan las reglas claras y sencillas, no es sensato tomar un p-valor como frontera universal y aplicarlo siempre con independencia del contexto en que esté situado. Fijar un valor frontera es lo mismo que decidir la probabilidad de equivocarnos si se rechaza la hipótesis nula, y la probabilidad de error que es razonable asumir depende, sin duda, de la situación en la que estemos y de las consecuencias de cometer el error.

Supongamos, por ejemplo, que un día por la mañana al salir de casa miramos como está el tiempo y consideramos que hay una probabilidad del 10% de que llueva ¿debemos volvernos a coger el paraguas? A nadie le parecerá temerario que sigamos y corramos un riesgo del 10% de que la lluvia nos pille desprotegidos. Si nos equivocamos no perdemos mucho (quizá nos mojemos un poco) y también hay que considerar que ir todo el día con el paraguas sin llover es bastante incomodo.

Otra situación. Vamos conduciendo por una carretera secundaria muy poco transitada. En un cambio de rasante, sin ninguna visibilidad de los coches que vienen de frente, vemos que hay un pequeño bache en el lugar por donde debemos pasar, aunque lo podríamos evitar colocándonos a la izquierda invadiendo el otro carril. Pero no lo haremos. La probabilidad de que por esa carretera tan poco transitada nos crucemos con un coche es pequeña, y que nos crucemos justo en el cambio de rasante es mucho menor. Pero no lo haremos porque aunque la probabilidad es muy pequeña, si se produce las consecuencias podrían ser muy graves. Además, pasar por el bache es sólo una ligera incomodidad. Es evidente que la probabilidad de error que estamos dispuestos a correr al tomar una decisión depende de las circunstancias y de lo que nos cuesta el error.

Por tanto, la respuesta a la pregunta planteada no es de carácter estadístico, sino que está relacionada con el contexto del problema que se está tratando. Cuando se realiza una prueba para analizar si un nuevo fármaco es mejor que el actual para curar una enfermedad, tomar como valor frontera 0,05 significa que corremos un riesgo del 5% de decir que es más eficaz cuando en realidad no es. ¿Qué implicaciones tiene esto? ¿Puede tener el nuevo tratamiento efectos secundarios perjudiciales? ¿Es mucho más caro que el tratamiento convencional? La respuesta a los interrogantes planteados son importantes para fijar el valor frontera más conveniente.

Pero también es verdad que en muchos casos se toma como referencia el valor de 0,05 sin entrar en más consideraciones. El porqué se utiliza 0,05 tiene que ver con los valores que figuran en las tablas. Cuando se empezaron a elaborar, con unos medios mucho más rudimentarios que ahora, solo se tabularon los valores correspondientes a algunas probabilidades –números fáciles como 0,001, 0,005; 0,01; 0,05, 0,10,…– y de entre los disponibles, se acostumbró a tomar el correspondiente a 0,05 como el más adecuado para separar lo habitual de lo raro. La ventaja del 0,05 es ser un valor redondo en nuestro sistema decimal. Si tuviéramos 6 dedos seguramente consideraríamos más natural tomar decisiones utilizando el 0,06 como valor frontera.

27 ¿Qué tipos de error se pueden cometer en un contraste de hipótesis?

Recordemos que se empieza suponiendo que las cosas son de una determinada forma, y que este planteamiento inicial, normalmente relacionado con no cambiar y seguir como estábamos, es lo que se denomina hipótesis nula. Esto será lo que creeremos a no ser que los datos lo contradigan, en cuyo caso consideraremos que se cumple la hipótesis alternativa (las hipótesis nula y alternativa se plantean de forma que sean excluyentes).

Por ejemplo, una línea de envasado de detergente llena los paquetes con un peso que sigue una distribución N(μ_X = 4 kg; σ_X=0,03 kg). Para comprobar que todo marcha correctamente cada cierto tiempo se toma una muestra, supongamos que de n=9 paquetes, y si su peso medio está fuera de un intervalo previamente fijado, los llamados límites de control, se considera que el proceso está descentrado. En este caso la hipótesis nula (lo normal, lo que se supone por defecto) es que el proceso está centrado ($H_0: \mu_X$ = 4 kg) y la alternativa que está descentrado ($H_1: \mu_X \neq$ 4 kg). El valor de la media de 9 paquetes, con el proceso centrado, se distribuye según N($\mu_{\bar{X}}$ = 4 kg; $\sigma_{\bar{X}} = \frac{0,03}{\sqrt{9}} = 0,01$ kg). Si los límites de control se fijan a ± 2 desviaciones tipo del valor nominal, la zona de rechazo será la que está fuera del intervalo 4,00 ± 0,02, tal como se indica en la Figura 27.1.

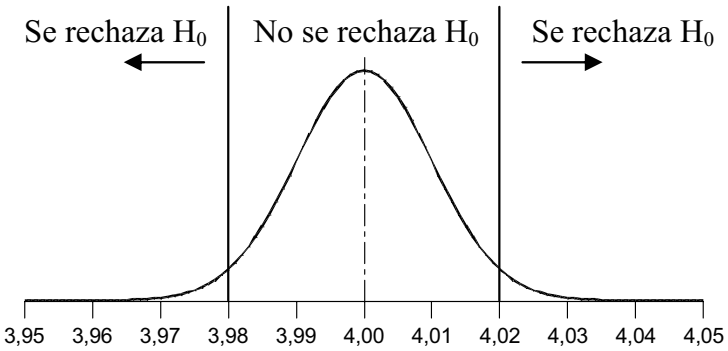


Figura 27.1. *Zonas de rechazo y no rechazo de la hipótesis nula si la población sigue una N(4; 0,03) y la decisión se toma de acuerdo a la media de una muestra de n=9, con límites a 2σ del valor nominal*

Esta forma de decidir implica que se pueden cometer 2 tipos de error:

1. Rechazar la hipótesis nula cuando en realidad es cierta. En nuestro caso esto significa que aunque el proceso esté centrado, es posible que la media de la muestra esté fuera del intervalo (3,98; 4,02). La probabilidad de que esto ocurra con nuestros números es del orden del 5% (exactamente 0,0455).

2. No rechazar la hipótesis nula cuando en realidad es falsa. Puede ser que la media muestral nos dé 4,01, en cuyo caso no rechazamos la hipótesis nula, pero el proceso podría estar centrado en 4,02 o en 4,03.

Al primer tipo de error se le llama error tipo I, y al segundo, error tipo II. La verdad es que, aparte de utilizar números romanos, hay poca originalidad en la denominación.

Podemos decidir la probabilidad de cometer el error tipo I fijando los límites[1] a partir de los cuales se rechaza la hipótesis nula. Por ejemplo, si los límites los fijamos a $\pm 3\sigma$, la probabilidad de rechazarla equivocadamente será del orden del 3 por mil (exactamente 0,0027). Si queremos que esta probabilidad sea del 1 por mil, los límites deberán estar en 3,967 y 4,033. A la máxima probabilidad de rechazar equivocadamente la hipótesis nula se la denomina α, y depende de dónde se coloca la frontera entre la zona de aceptación y la de rechazo[2].

Veamos ahora que pasa con el error tipo II. La probabilidad de cometer este tipo de error no se puede calcular sin fijar un valor concreto para la hipótesis alternativa. Por ejemplo, si el proceso se descentra y pasa a llenar los paquetes con un peso medio de 4,03, la probabilidad de que la media de una muestra de 9 paquetes esté dentro de los límites de control es de 0,1587. Luego existe una probabilidad del orden del 16% de que consideremos que el proceso está centrado en su valor nominal, cuando en realidad lo está en 4,03. A esta probabilidad de cometer el error tipo II se la denomina β.

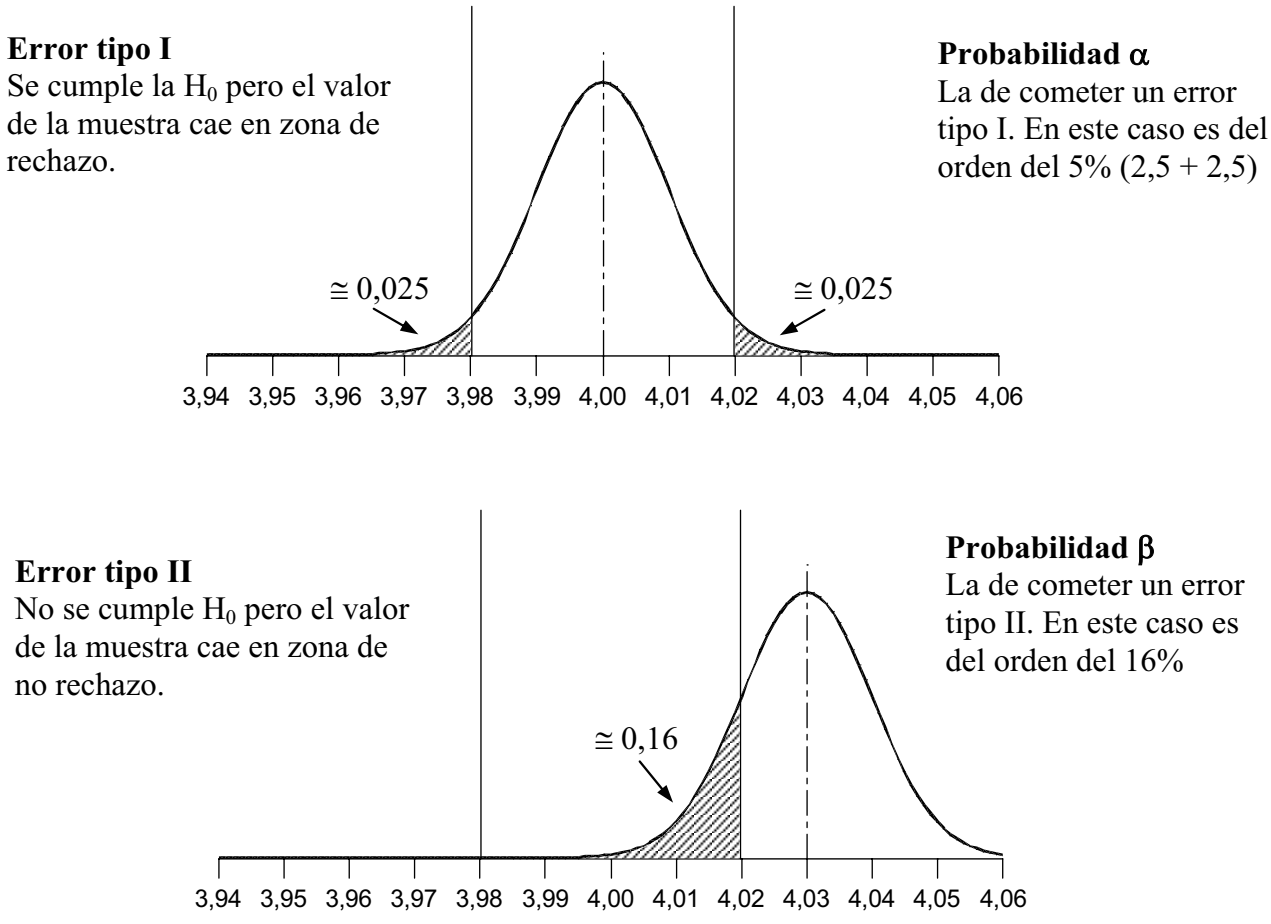


Figura 27.2. *Probabilidad de cometer el error tipo I con los límites en 4.00±0,02 y probabilidad de cometer el tipo II si el proceso se centra en 4,03*

[1] Si la hipótesis alternativa fuera del tipo "mayor que" o "menor que", tendríamos un solo límite.

[2] Esta probabilidad no es el p-valor que se obtiene en el contraste de hipótesis. El p-valor varía de una muestra a otra y la probabilidad de rechazar equivocadamente la hipótesis nula es una sola, que depende de la regla de decisión establecida. Si no cambia la regla de decisión, no cambiará esta probabilidad de equivocarnos.

Observe que mientras la probabilidad α depende solo de la regla de decisión (situación de los límites de control), la β depende además del valor que tome la media. Al valor $1-\beta$ se le denomina potencia de la prueba para detectar una media de 4,03. Si calculamos $1-\beta$ para los distintos valores posibles de la media μ, definidos en la hipótesis alternativa, a dicha curva se la conoce como "curva de potencia de la prueba".

Tabla 27.1. *Valores de la probabilidad β y de la potencia $(1-\beta)$ para distintos valores de μ*

	μ	β	$1-\beta$
	3,94	0,000032	0,999968
	3,95	0,001350	0,998650
	3,96	0,022750	0,977250
	3,97	0,158655	0,841345
	3,98	0,499968	0,500032
	3,99	0,839995	0,160005
	4,00	0,954500	0,045500
	4,01	0,839995	0,160005
Valor calculado	4,02	0,499968	0,500032
en el ejemplo →	4,03	0,158655	0,841345
	4,04	0,022750	0,977250
	4,05	0,001350	0,998650
	4,06	0,000032	0,999968

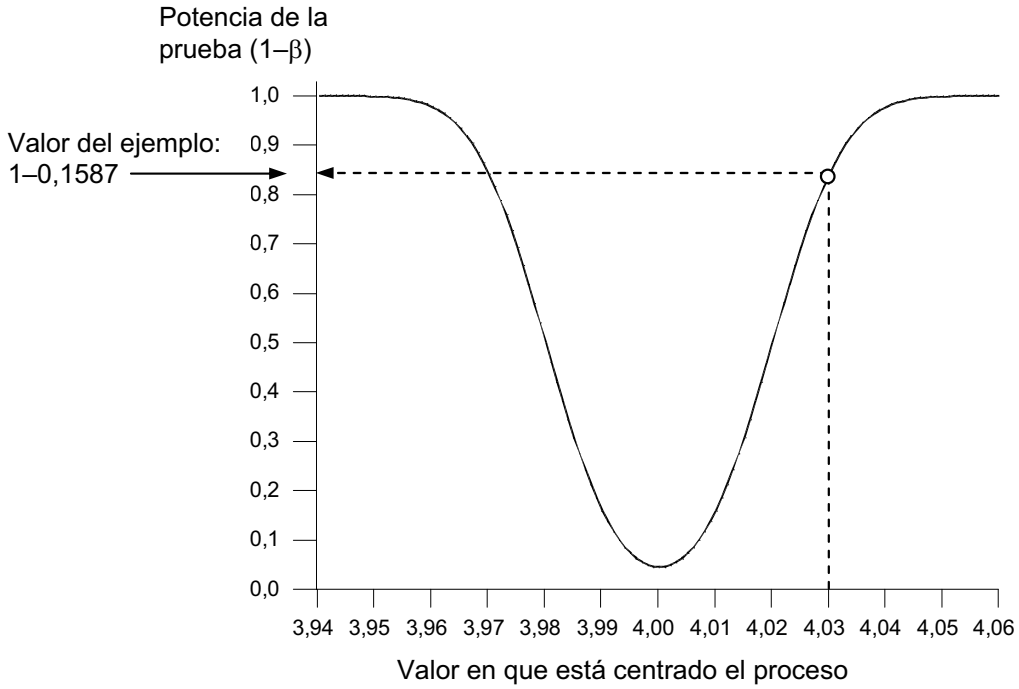


Figura 27.3. *Curva de potencia de nuestra prueba*

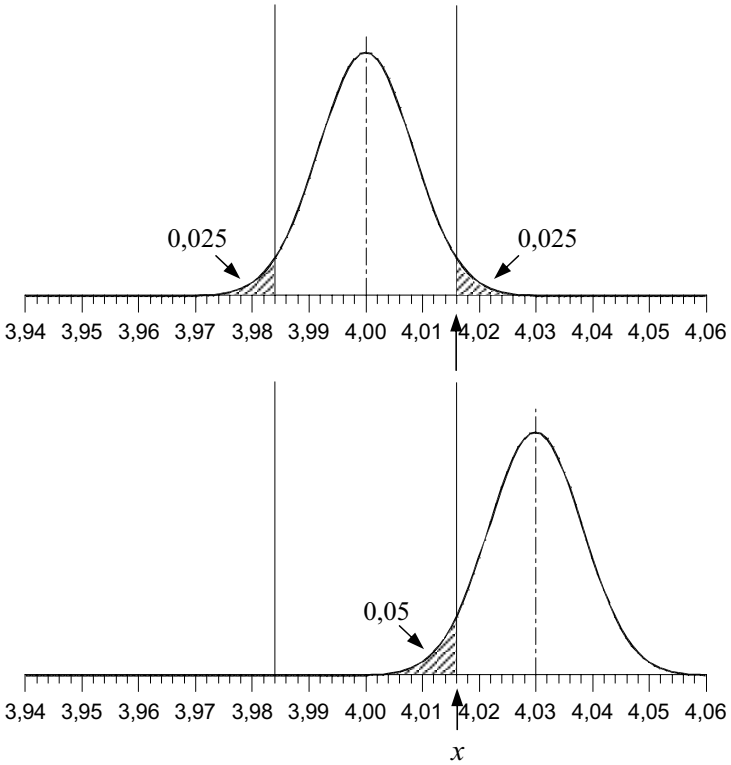
¿Cómo disminuir la probabilidad β? Una opción es aumentar α. Si en nuestro caso los límites de control los ponemos a $\pm 1\sigma$, la probabilidad α pasará a ser del 32%, pero para $\mu = 4{,}03$ en vez de ser del 16%, lo será del 2,5%.

¿Y si queremos disminuir α? Entonces aumenta β. Si ponemos los límites a ±3σ, α pasa a ser del 3 por mil, pero β, para μ=4,03, pasa a ser del 50%.
¿Es posible disminuir α y β a la vez? Sí, aumentando el tamaño de la muestra. Por ejemplo, si el tamaño de la muestra en vez de ser 9 fuera 25, la desviación tipo de la media pasaría a ser $\frac{0,03}{\sqrt{25}} = 0,006$, y una probabilidad α de 0,05 significa poner los límites a ±2σ, es decir a 4,00±0,012. En este caso, si el proceso se descentra pasando a tener un valor medio de 4,03, la probabilidad β será prácticamente cero.

Del trío α, β y *n*, dados 2 cualesquiera, se puede deducir el tercero. Por ejemplo, si en nuestro caso queremos una probabilidad α del 5% y una β también de 5%, cuando la media se coloque en 4,03, el tamaño de la muestra debe ser[3] *n*=13.

En definitiva, para un esfuerzo dado (esfuerzo = tamaño de muestra) si queremos reducir α debe ser a costa de β y viceversa. Si se quieren reducir las 2 probabilidades a la vez no queda más remedio que hacer más esfuerzo (más muestra). En este contexto también vale aquello de que "el que algo quiere algo le cuesta".

[3] $z_{0,025} = 1,96 \rightarrow 1,96 = \frac{x-4,00}{\frac{0,03}{\sqrt{n}}}$; $z_{0,05} = 1,65 \rightarrow -1,65 = \frac{x-4,03}{\frac{0,03}{\sqrt{n}}}$; de donde x = 4,016 y n = 13



28 ¿Es correcto multiplicar por 2 el área de cola en los tests de igualdad de varianzas cuando H_1 es del tipo "distinto de"?

En este tipo de test el estadístico de prueba es el cociente de las varianzas muestrales, y la distribución de referencia es la F de Snedecor con los grados de libertad de las varianzas que hemos colocado en el numerador y el denominador (por este orden).

Como la distribución F de Snedecor no es simétrica, parece que para hallar el área de las 2 colas, cuando la hipótesis alternativa es del tipo "distinto de", no se puede multiplicar por 2 el área de una, tal como hacemos cuando la distribución es simétrica como la Normal o la *t* de Student.

Pues bien, sí se puede multiplicar por 2 el área de cola también en este caso. Sean s_A^2 y s_B^2 las 2 varianzas que queremos comparar, y supongamos que $s_A^2 > s_B^2$. El estadístico de prueba es el cociente de las 2 varianzas, y podemos elegir tanto s_A^2/s_B^2 , como al revés, es decir, su inversa s_B^2/s_A^2 .

Si elegimos s_A^2/s_B^2 , como el numerador es mayor que el denominador, tendremos un cociente mayor que 1, y será tanto mayor cuanto mayor sea la diferencia entre varianzas. En este caso tendremos que mirar el área de cola hacia la derecha en una distribución F de Snedecor con ν_A y ν_B grados de libertad.

Si por el contrario, elegimos s_B^2/s_A^2 al ser $s_A^2 > s_B^2$ el cociente será menor que 1, tanto menor cuanto mayor sea la diferencia entre varianzas. En este caso, lo que interesa es la cola hacia la izquierda, pero no en la misma distribución que en el caso anterior, porque ahora se ha cambiado el orden de los grados de libertad.

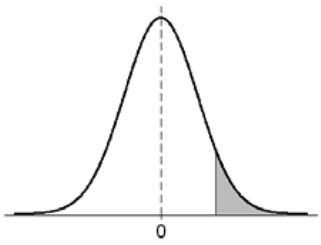
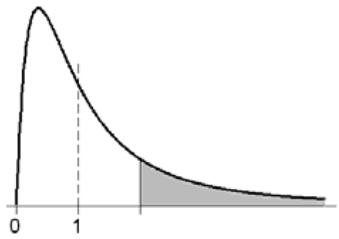
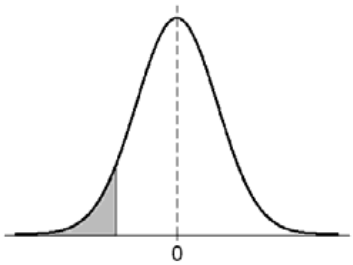
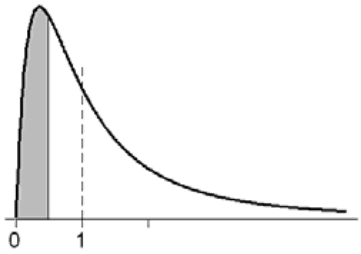
Para comparar estas 2 áreas utilizaremos la siguiente propiedad de la distribución F de Snedecor:

$$F_{\nu_A,\nu_B}(\alpha) = \frac{1}{F_{\nu_B,\nu_A}(1-\alpha)}$$

El valor entre paréntesis es el área de cola que deja hacia la derecha una distribución con los grados de libertad que se indican. De aquí se deduce que las 2 áreas que hemos hallado son idénticas y que basta con calcular solo una y multiplicarla por 2 (ver Figura 28.1). Como no es necesario mirar las 2 áreas de cola, las tablas que vienen en los libros solo dan colas hacia la derecha y el estadístico que se usa es el cociente de la varianza mayor partido por la menor (cociente mayor que 1, área de cola hacia la derecha).

Figura 28.1. *Determinación de las 2 áreas de cola en un test de igualdad de varianzas*

Tabla 28.1. *Analogía entre el test de comparación de medias y el de comparación de varianzas (H_1 tipo ≠)*

	Comparación de medias	Comparación de varianzas
Planteamiento	H_0: $\mu_A = \mu_B$ H_1: $\mu_A \neq \mu_B$	H_0: $\sigma_A^2 = \sigma_B^2$ H_1: $\sigma_A^2 \neq \sigma_B^2$
Tamaño de las muestras	n_A, n_B	n_A, n_B
Supongamos que:	$\overline{y}_A > \overline{y}_B$	$s_A^2 > s_B^2$
Estadístico de prueba usado normalmente	$\dfrac{\overline{y}_A - \overline{y}_B}{s\sqrt{\dfrac{1}{n_A}+\dfrac{1}{n_B}}}$	$\dfrac{s_A^2}{s_B^2}$
Característica de este estadístico	Es positivo	Es mayor que 1. Está tabulado
Distribución de referencia	t Student con $n_A + n_B - 2$ g.l.	F Snedecor con n_A–1 y n_B–1 g.l.
Si los estadísticos muestrales son iguales, el estadístico de prueba es	0	1
Área de cola 1		
Otro posible estadístico de prueba	$\dfrac{\overline{y}_B - \overline{y}_A}{s\sqrt{\dfrac{1}{n_A}+\dfrac{1}{n_B}}}$	$\dfrac{s_B^2}{s_A^2}$
Característica de este estadístico	Es negativo	Es menor que 1 No está en las tablas
Distribución de referencia	t Student con $n_A + n_B - 2$ g.l.	F Snedecor con n_B–1 y n_A–1 g.l. (ojo, no es la misma que antes)
Área de cola 2		
p-valor	Área cola 1 + Área cola 2	Área cola 1 + Área cola 2
	En ambos casos, el área 1 y el área 2 son iguales. Basta calcular una y multiplicarla por 2	

29 ¿Por qué respecto a la hipótesis nula se habla de "no rechazo" y no de "aceptación"?

Esta es una característica general de la búsqueda del conocimiento científico. Ilustraremos la situación con un ejemplo.

En un juego se trata de descubrir un animal oculto y los participantes (los científicos) deben hacer preguntas sobre características de dicho animal con el propósito de descubrirlo. Uno de ellos tiene como hipótesis (nula) que el animal es una paloma y para contrastar su hipótesis pregunta: ¿el animal en cuestión tiene alas? Se le responde que sí, que efectivamente el animal tiene alas.

¿Qué conclusión puede sacarse hasta el momento? ¿Podría decirse entonces que su hipótesis es cierta y que el animal en cuestión es necesariamente una paloma? Sabemos que la respuesta es no. La evidencia (tener alas) es compatible con su hipótesis, pero ello no significa que sea verdadera, pues existen muchas otras hipótesis que son igualmente compatibles con dicha evidencia, por ejemplo, que el animal sea una mariposa, o que sea un murciélago.

El científico hace una nueva pregunta (observación): ¿el animal es vertebrado? Y como resultado de la observación se le responde que no. ¿Qué pasa con su hipótesis ahora? ¿El animal podría ser una paloma? No. Ahora rechazamos la hipótesis de forma contundente para replantearla de tal manera que sea compatible con los nuevos hechos: tiene alas y no es vertebrado. La nueva hipótesis de trabajo puede ser: el animal es un mosquito.

Todas las verdades en la ciencia son de carácter transitorio, de forma que una afirmación sobre la naturaleza es verdadera porque no ha podido demostrarse que es falsa. Sin embargo, pueden existir muchas hipótesis compatibles con los hechos (no solo la nuestra), por esta razón el no rechazo de una hipótesis no implica su veracidad. No ocurre lo mismo cuando los hechos contradicen la hipótesis, así, si el animal no es vertebrado, estamos muy seguros de que no es una paloma.

Volviendo a nuestro tema, al contrastar la hipótesis nula de que la media poblacional es $\mu=10$ frente a la alternativa $\mu\neq10$, suponiendo que la desviación tipo de la población es $\sigma=6$, si se obtiene que la media de una muestra de 9 observaciones es $\bar{x}=25$, claramente podemos rechazar la hipótesis nula, pero si sale $\bar{x}=11$ no la podemos rechazar, aunque tampoco afirmar que es cierta, porque nuestros datos también serían coherentes con hipótesis del tipo $\mu=10{,}5$ o $\mu=11$.

30 ¿Es lo mismo diferencia significativa que diferencia importante?

No es lo mismo. Una diferencia altamente significativa puede ser irrelevante a efectos prácticos, y también puede ocurrir que una diferencia importante no sea significativa.

En los procesos de soldadura por puntos se hace pasar una gran cantidad de corriente eléctrica (miles de amperios) durante un corto espacio de tiempo (milisegundos) pero suficiente para que las 2 superficies metálicas se fundan en los puntos de contacto con los electrodos por el calor producido por el efecto Joule. Al fundirse se mezclan los materiales de las 2 chapas y la unión que forman al solidificarse es lo que llamamos punto de soldadura.

Un problema que se puede presentar en estos procesos está relacionado con el tratamiento superficial que se da a las chapas para que no se oxiden durante el periodo en que están almacenadas o manipulándose en el taller. Este tratamiento suele consistir en un recubrimiento con cierto producto (le llamaremos pintura) que no es tan buen conductor como el metal y dificulta el paso de la corriente, lo que puede repercutir en las características del punto de soldadura.

Se decide comparar la pintura habitual, A, con otra alternativa, B, que aseguran es más conductora y permite mejores soldaduras. Para realizar la comparación se llevan a cabo un cierto número de soldaduras en chapas pintadas con A, y otras en chapas pintadas con B. Lo que se mide es la resistencia a la cizalladura. Vamos a analizar cuáles serían nuestras conclusiones en 2 situaciones con distintos resultados.

	Situación 1		Situación 2	
	Pintura A	Pintura B	Pintura A	Pintura B
Tamaño de las muestras	$n_A = 100$	$n_B = 100$	$n_A = 10$	$n_B = 10$
Media	$\overline{y}_A = 1750$ kg	$\overline{y}_B = 1752$ kg	$\overline{y}_A = 1750$ kg	$\overline{y}_B = 1850$ kg
Desviación tipo	$s_A = 3{,}70$ kg	$s_B = 3{,}69$ kg	$s_A = 169$ kg	$s_B = 193$ kg
$\overline{y}_B - \overline{y}_A$	2 kg		100 kg	
Estimador conjunto de σ: $s = \sqrt{\frac{(n_A-1)s_A^2 + (n_B-1)s_B^2}{n_A+n_B-2}}$	$s_1 = 3{,}69$ kg		$s_2 = 181$ kg	
Estadístico de prueba[1]: $t = \frac{\overline{y}_B - \overline{y}_A}{s\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}}$	$t_1 = 3{,}83$		$t_2 = 1{,}23$	
Distribución de referencia	t-Student con 198 g.l.		t-Student con 18 g.l.	
p-valor (H_1: $\mu_B > \mu_A$)	0,000		0,117	

[1] La justificación de las fórmulas que hemos utilizado para el estimador conjunto de σ y para el estadístico de prueba pueden encontrarse, por ejemplo, en *Fundamentos de estadística* de Daniel Peña, Alianza Editorial, 2001.

Si los resultados son como los reflejados en la situación 1, la diferencia es claramente significativa, pero 2 kg sobre 1.750 seguramente es una diferencia poco importante, y si cambiar de pintura entraña alguna dificultad o algún riesgo, por pequeño que sea, lo más probable es que no se cambie.

En la situación 2 la diferencia son 100 kg, lo cual ya puede ser una diferencia importante, al menos por comparación con el caso anterior. Pero ahora la diferencia no es significativa, es decir, puede ser debida al azar y no estaría justificado cambiar de pintura a causa de estos resultados, pues si se repite el experimento es posible que resulte una diferencia de –100 kg.

Obsérvese que hay 2 características que tienen distinto orden de magnitud en ambas situaciones: el tamaño de las muestras y las desviaciones tipo. En la situación 1 (diferencia significativa) el tamaño de las muestras es grande, n=100, y la desviación tipo pequeña, s_A=3,70 y s_B=3,69, frente a la situación 2 (diferencia no significativa) en que el tamaño de las muestras es 10 y las desviaciones tipo son 169 y 193 kg. Tamaños de muestra grandes, con poca variabilidad dentro de las muestras, permiten detectar diferencias muy pequeñas y quizá irrelevantes a efectos prácticos. Sin embargo, si los tamaños de muestra son pequeños y la variabilidad grande, no se ponen de manifiesto las diferencias aunque estas existan y sean importantes.

Por tanto, hay que andarse con cuidado con las diferencias significativas, porque el hecho de que lo sean no implica que también sean importantes. Y no hay que confundir diferencia no significativa con no existencia de diferencias. Es posible que el pequeño tamaño de muestra y/o la variabilidad que existe en los datos no nos deje ver una diferencia que sí existe, y que además puede ser importante.

Comparación de tratamientos

31 ¿Cómo elegir la hipótesis alternativa que conviene plantear?

La selección de la hipótesis alternativa está relacionada con el planteamiento del problema y con la zona en que queremos situar el riesgo de equivocarnos al rechazar la hipótesis nula. Lo veremos a través de la reflexión sobre 3 situaciones distintas.

Situación 1: Fertilizante enriquecido para una plantación de tomates

Una explotación agraria es informada de que añadiendo un cierto complejo mineral al fertilizante que usa en sus plantaciones de tomates, se obtiene una mayor cosecha. Para comprobar si esta afirmación es cierta se toman 20 plantas, en 10 de ellas se utiliza sólo el fertilizante habitual (grupo de control A) y en las otras 10 se añade el complemento mineral (grupo tratado B). Se quiere contrastar la hipótesis nula de que el peso medio de tomates por planta es igual en ambos casos. ¿Cuál debe ser la alternativa?

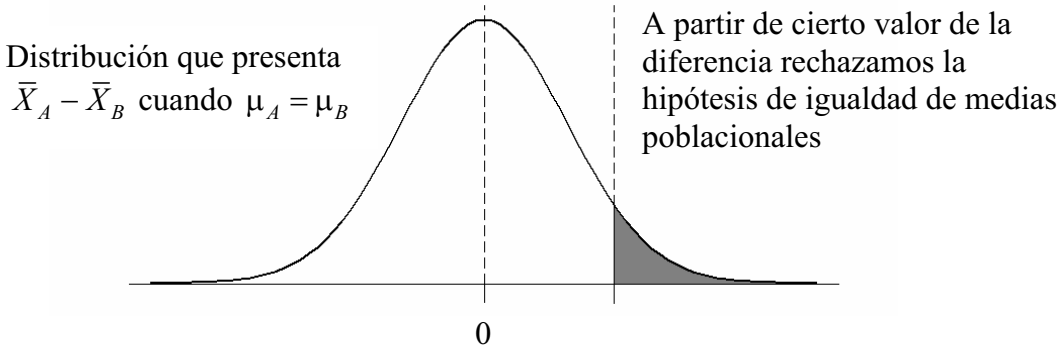


Figura.31.1. *Sólo rechazamos H_0 si la media del grupo tratado es "mucho" mayor*

Como nuestro interés está en saber si el complemento mineral es eficaz, sólo rechazaremos la hipótesis nula si la media del grupo tratado es mayor (en la magnitud requerida) que la media del grupo de control. Naturalmente no podremos decir que el complejo mineral es eficaz si la media del grupo tratado es menor. Por tanto, el riesgo de equivocarnos lo situamos sólo hacia los valores mayores. La hipótesis alternativa debe ser: H_1: $\mu_A < \mu_B$

Situación 2: ¿Tienen los sacos de fertilizante un peso menor al establecido?

La explotación agrícola compra sacos de 50 kg de fertilizante, y el proveedor asegura que el peso medio de los sacos es efectivamente de 50 kg, con una desviación tipo de 0,5 kg. Como el comprador sospecha que los sacos contienen en promedio menos de 50 kg, decide tomar una muestra de sacos y contrastar la hipótesis nula de que el peso medio es el que dice el proveedor, H_0: $\mu = 50$, ¿frente a qué alternativa?

Para realizar la prueba podemos tomar 4 sacos, y comparar su peso medio con la distribución a que pertenecería si μ=50. Lo que nos preocupa es que el peso sea menor y por tanto, sólo rechazaremos la hipótesis nula si la media de los 4 sacos es suficientemente menor que el valor nominal. Es decir, en este caso la hipótesis alternativa será: H_1: $\mu < 50$, y con un nivel de significación α=0,05, "suficientemente menor" significa menor que

$$50 - z_{0,05}\frac{\sigma}{\sqrt{4}} = 50 - 1,645 \cdot \frac{0,5}{2} = 49,59$$

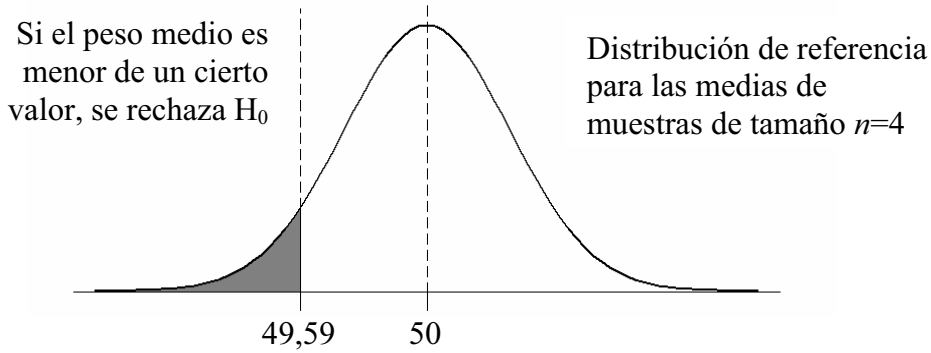


Figura 31.2. *Zona de rechazo de H_0 si lo que nos preocupa es que los sacos pesen menos de 50 kg*

Obsérvese que este planteamiento beneficia al vendedor, ya que la carga de la prueba se le pone al comprador. El proceso se supone centrado "a no ser que se demuestre lo contrario", y el riesgo de equivocarnos en contra del vendedor (que el promedio sea de 50 kg y digamos que es menor) está fijado de antemano en un valor bajo (riesgo α, en nuestro caso 0,05), pero nada se dice sobre el riesgo de que el promedio sea, por ejemplo, 49,5 kg y que demos el proceso por bueno. Este último riesgo (riego del comprador, probabilidad β) con los números de nuestro ejemplo es igual al 36% (nada despreciable).

Situación 3: ¿Está el peso de los sacos fuera de tolerancias?

El productor llena los sacos en una línea automática, y cada cierto tiempo pesa una muestra para verificar que el peso medio está en 50 kg. Al fabricante no le interesa dar más producto (pierde dinero) ni menos (da mala calidad). Su hipótesis nula es que el proceso se mantiene centrado. ¿Cuál debe ser la alternativa?

Tal como está planteado, al fabricante le preocupa lo mismo que el proceso se descentre hacia arriba o hacia abajo, así que deberá estar atento hacia los dos lados. Si su plan de control consiste en tomar muestras de 4 sacos y si acepta tener un 5% de falsas alarmas, deberá repartir este riesgo entre valores por defecto (2,5%) y valores por exceso (2,5%). En este caso los valores críticos son: $50 \pm z_{0,025} \frac{\sigma}{\sqrt{4}}$, lo que resulta igual a 49,51 y 50,49.

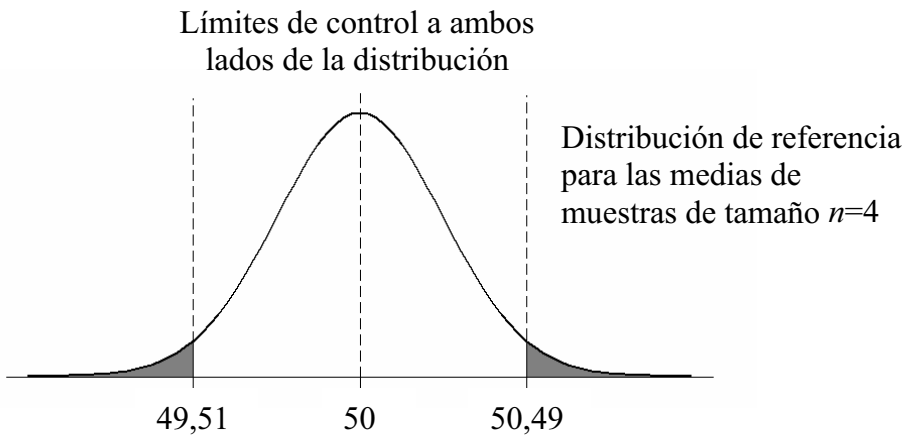


Figura 31.3. *Zona de rechazo de H_0 si lo que nos preocupa es que el peso de los sacos se desvíe del valor nominal tanto por exceso como por defecto*

Evidentemente en este caso la hipótesis alternativa debe ser H_1: $\mu \neq 50$.

32 ¿Si la hipótesis alternativa es del tipo "mayor que" o "menor que", ¿cómo se sabe hacia qué lado hay que mirar el área de cola?

Cuando se plantean dudas de este tipo siempre es mejor encontrar la respuesta razonando, ya que echar mano de recetas, especialmente si se acepta el resultado obtenido sin ningún espíritu crítico, puede llevar a cometer errores de mucho bulto. Supongamos que el planteamiento del contraste es:

$$H_0: \mu_A = \mu_B$$
$$H_1: \mu_A < \mu_B$$

Con los resultados de las medias muestrales estaremos en uno de los siguientes casos:

a) $\bar{y}_A > \bar{y}_B$: Estamos de suerte (en cuanto a facilidad de cálculo). No hace falta que calculemos nada. Seguro que no se puede rechazar la hipótesis nula. Por ejemplo, si se analiza la eficacia de un complemento mineral como fertilizante para una plantación de tomates, y resulta que las plantas con que se ha utilizado este complemento dan una producción menor, es evidente que no podremos rechazar la hipótesis nula y concluir que el complemento mineral aumenta la cosecha.

b) $\bar{y}_A < \bar{y}_B$: Seguramente este es el resultado que esperábamos, y ahora lo que toca es analizar si la diferencia observada puede considerarse estadísticamente significativa. En este ejemplo del complemento para el abono podremos elegir como estadísticos de prueba[1]:

$$t_1 = \frac{\bar{y}_A - \bar{y}_B}{s\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \qquad t_2 = \frac{\bar{y}_B - \bar{y}_A}{s\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

La única diferencia entre uno y otro es que t_1 será negativo y t_2 positivo (exactamente $t_1 = -t_2$). Si se tiene el valor positivo habrá que mirar el área de cola hacia la derecha para ver qué tan lejos está el valor obtenido de la distribución de referencia, y si es negativa habrá que mirar la cola hacia la izquierda. Obsérvese que si se hace al revés, cuanto más raro sea el valor (más alejado quede del centro de la distribución) mayor sería el área de cola.

[1] Entendemos que es razonable suponer que la población es Normal o la muestra es grande, y que las varianzas de las 2 poblaciones son iguales.

Aunque en esencia es exactamente lo mismo, gusta más trabajar con valores positivos. Es más cómodo y además son los que encontramos en casi todas las tablas.

c) $\overline{y}_A = \overline{y}_B$ Si esto ocurre, no hace falta calcular nada, sea cual sea la hipótesis alternativa. (¡Cómo vamos a demostrar que una diferencia es significativa si no existe tal diferencia!).

Siempre hay que comprobar la coherencia del resultado obtenido con el planteamiento realizado. Por ejemplo, el área de cola no puede ser mayor de 0,5.

La Tabla 32.1 resume las acciones a llevar a cabo según el planteamiento del contraste y el sentido de la desigualdad entre las medias muestrales.

Tabla 32.1. *Acciones a llevar a cabo según el planteamiento del contraste y el sentido de la desigualdad entre las medias muestrales*

<table>
<tr><th rowspan="2">Resultado obtenido</th><th colspan="3">PLANTEAMIENTO DEL CONTRASTE</th></tr>
<tr><th>H_0: $\mu_A = \mu_B$
H_1: $\mu_A < \mu_B$</th><th>H_0: $\mu_A = \mu_B$
H_1: $\mu_A > \mu_B$</th><th>H_0: $\mu_A = \mu_B$
H_1: $\mu_A \neq \mu_B$</th></tr>
<tr><td>$\overline{y}_A < \overline{y}_B$</td><td>Resultado esperado. Se tratará de analizar mediante el procedimiento adecuado si la diferencia obtenida es o no estadísticamente significativa.</td><td>No hace falta que realicemos ningún cálculo. Con el resultado obtenido es obvio que no podemos rechazar H_0 para quedarnos con H_1.</td><td rowspan="2">Es necesario analizar si la diferencia obtenida es estadísticamente significativa.</td></tr>
<tr><td>$\overline{y}_A > \overline{y}_B$</td><td>No hace falta que realicemos ningún cálculo. Con el resultado obtenido es obvio que no podemos rechazar H_0 para quedarnos con H_1.</td><td>Resultado esperado. Se tratará de analizar mediante el procedimiento adecuado si la diferencia obtenida es o no estadísticamente significativa.</td></tr>
<tr><td>$\overline{y}_A = \overline{y}_B$</td><td colspan="3">En este caso, que prácticamente no se dará (sería una casualidad), obviamente no podrá rechazarse la hipótesis nula sea cual fuere la alternativa.</td></tr>
</table>

33 ¿Por qué el análisis de la varianza se llama así, cuando en realidad se trata de una técnica para comparar medias y no varianzas?

No es por ganas de confundir a los estudiantes. Se verá claro con un ejemplo. Supongamos que se desean comparar las medias de las siguientes muestras:

A:	13,8	13,8	13,7	15,2	14,1	14,3	14,1	13,5	13,4	14,0
B:	16,3	15,4	15,2	15,1	15,6	16,0	14,9	16,5	16,1	16,0
C:	13,3	13,3	14,1	13,6	13,4	13,7	13,9	14,1	13,3	14,6

Una buena idea es empezar con el análisis gráfico de los datos. Los diagramas de puntos presentan el aspecto que se indica en la Figura 33.1.

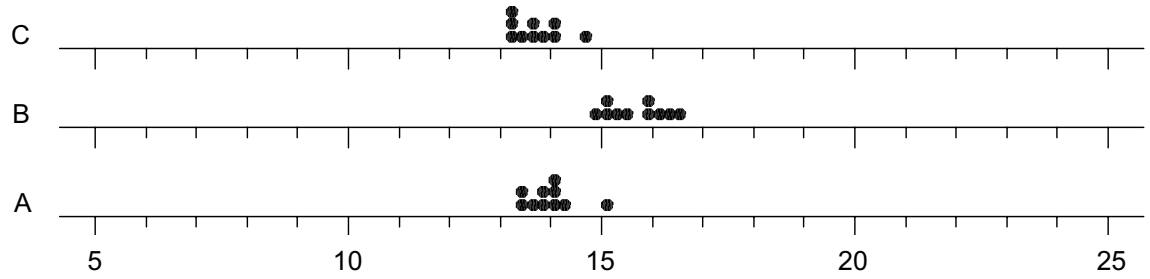


Figura 33.1

¿Qué impresión nos da este gráfico? Estará de acuerdo que parece claro afirmar que B da un nivel de respuesta mayor que A o C, no apreciándose diferencias entre estos 2 últimos.

Veamos ahora otra situación. Supongamos que los valores obtenidos son:

D:	11,4	19,5	10,7	14,2	14,8	8,0	21,2	8,7	21,8	9,2
E:	14,4	15,1	12,2	20,5	10,5	19,9	23,0	14,3	10,8	16,4
F:	14,9	16,4	14,2	7,0	11,8	7,7	17,4	10,4	20,3	17,3

Los diagramas de puntos ahora tienen el aspecto que indica la Figura 33.2.

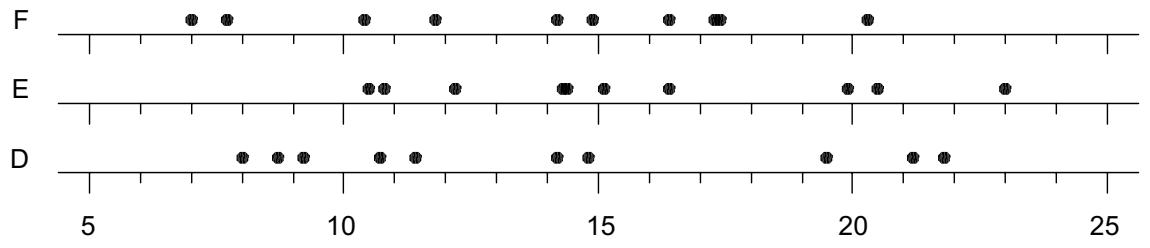


Figura 33.2

¿Qué opina sobre las diferencias de medias en este caso? No se ve que las diferencias sean significativas. Las 3 muestras pueden provenir perfectamente de la misma población.

Veamos ahora cuáles son las medias en ambos casos:

Caso 1		Caso 2	
Muestra	Media	Muestra	Media
A	14,0	D	14,0
B	15,7	E	15,7
C	13,7	F	13,7

¡Las diferencias de medias son iguales en ambos casos! Entonces, ¿por qué las conclusiones han sido distintas?

Porque, de una forma intuitiva, hemos comparado la variabilidad de los datos en cada muestra con las diferencias (es decir, también variabilidad) que presentan sus medias. En el primer caso, si las 3 muestras provienen de la misma población, las diferencias entre sus medias son mayores de lo que cabe esperar a partir de la variabilidad dentro de cada muestra. En el segundo caso, la variabilidad dentro de las muestras es mucho mayor y por eso no son sorprendentes las diferencias entre las medias suponiendo que las 3 muestras provengan de una misma población.

Por tanto, en ambos casos hemos llegado a nuestras conclusiones sobre la igualdad de medias analizando variabilidades. Cuando la comparación se realiza analíticamente (en vez de gráficamente, como hemos hecho nosotros) las variabilidades se miden a través de las varianzas. Lo que se hace es comparar varianzas. Por eso la técnica se llama análisis de la varianza.

34 ¿Por qué para comparar *k* tratamientos se utiliza la técnica de análisis de la varianza, en vez del ya conocido test de la *t* de Student aplicándolo a todas las parejas que se pueden formar con *k* tratamientos?

Utilizar a fondo las técnicas disponibles es, desde luego, una buena idea, pero que no sirve para el caso que nos ocupa. No es lo mismo comparar todas las parejas de tratamientos que aplicar la técnica de análisis de la varianza. Vamos a verlo.

Supongamos, por concretar, que tenemos 5 tratamientos: A, B, C, D y E. Podemos calcular las medias: $\overline{y}_A$, $\overline{y}_B$, ... y plantear para cada una de las 10 parejas que se pueden formar la hipótesis nula de que sus medias poblacionales son iguales frente a la alternativa de que son distintas.

$$\begin{aligned}
&\text{Pareja 1:} && \mu_A = \mu_B \text{ frente a } \mu_A \neq \mu_B \\
&\text{Pareja 2:} && \mu_A = \mu_C \text{ frente a } \mu_A \neq \mu_C \\
& && \vdots \\
&\text{Pareja 10:} && \mu_D = \mu_E \text{ frente a } \mu_D \neq \mu_E
\end{aligned}$$

Estos 10 contrastes se pueden abordar construyendo los 10 intervalos de confianza para las diferencias de las medias poblacionales. Si no existen diferencias entre los tratamientos la probabilidad de que un intervalo incluya el cero y por tanto nos conduzca a la conclusión correcta, es igual a su nivel de confianza, supongamos –por ejemplo– del 95%.

¿Qué pasa con el análisis conjunto? Al ser iguales todos los tratamientos, solo llegaremos a la conclusión correcta en el caso de no rechazar ninguna de las hipótesis nulas. Si fueran independientes, la probabilidad de que esto ocurra será de $0{,}95^{10} = 0{,}60$.

El riesgo de rechazar "injustamente" la hipótesis nula aumenta al aumentar el número de tratamientos que se comparan, de forma que los riesgos no son los que parecen a primera vista. Si las pruebas fueran independientes, para comparar *k* parejas con un nivel de confianza *c* global para toda la prueba, bastaría con calcular los intervalos individuales con un nivel de confianza $\sqrt[k]{c}$. Pero no se puede contar con la independencia, ya que si un par de tratamientos presentan una diferencia significativa porque uno de ellos tiene la media anormalmente alta, esto aumenta la probabilidad de que otras diferencias aparezcan también como significativas.

El análisis de la varianza resuelve este problema al plantear un contraste de hipótesis conjunto sobre la igualdad de todas las medias frente a la alternativa de que alguna es distinta. Su punto débil es que cuando se rechaza la hipótesis nula no informa sobre cuál o cuáles tratamientos deben ser considerados distintos.

Una buena forma de actuar es aplicar la técnica de análisis de la varianza y, si se rechaza la hipótesis de igualdad de medias, aplicar una técnica para la jerarquización de las medias como el método de Tukey, que puede verse en el texto de Box, Hunter y Hunter.

Correlación y Regresión

35 ¿Por qué cuando se ajusta una nube de puntos a una ecuación de regresión, se utiliza siempre el criterio de minimizar la suma de los cuadrados de los residuos, y no otros como minimizar la suma de su valor absoluto?

Cuando se cumplen las llamadas "hipótesis del modelo"[1], el método de los mínimos cuadrados reúne un conjunto de propiedades que le dan importantes ventajas respecto a otros criterios de ajuste[2].

En concreto, minimizar la suma del valor absoluto de los residuos puede conducir a ajustes no adecuados. Por ejemplo, sean los puntos:

X	Y
4	4
10	10
16	10

Si se ajustan a una recta con el criterio de los minimizar la suma de los cuadrados de los residuos (puede usarse cualquier calculadora o paquete de software que lo dé directamente), se obtiene la ecuación: $Y = 3 + 0{,}5\,X$.

Esta ecuación representa una recta que se ajusta de forma muy razonable a nuestros puntos. Seguramente a mano trazaríamos una muy parecida. La suma de los cuadrados de los residuos es $\sum e_i^2 = 6$ (el mínimo valor posible, dado el criterio de cálculo) y la suma de los residuos en valor absoluto resulta $\sum |e_i| = 4$, tal como se puede comprobar en la Figura 35.1.

En la Figura 35.2 se representa otro ajuste. En este caso con $\sum e_i^2 = 9$ y $\sum |e_i| = 3$, y aunque ha disminuido la suma de los residuos en valor absoluto, el ajuste no es el que parece más razonable.

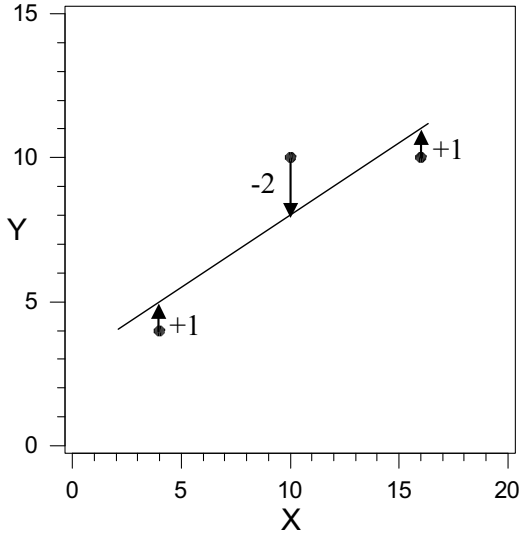


Figura 35.1. *Ajuste por el método de los mínimos cuadrados. Suma de residuos en valor absoluto = 4*

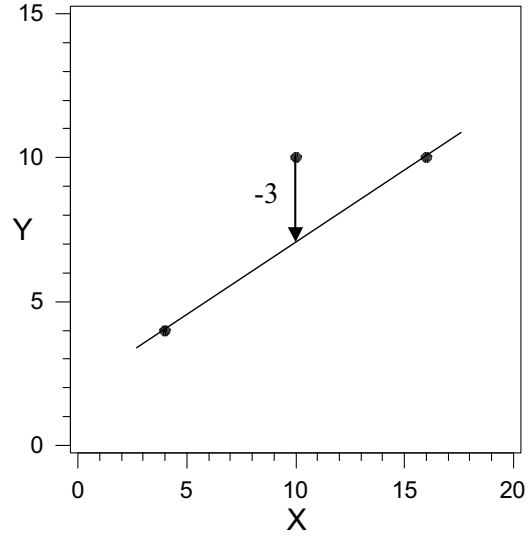


Figura 35.2. *Ajuste que proporciona una suma de residuos en valor absoluto igual a 3*

[1] Básicamente se considera que los errores aleatorios son independientes unos de otros y se distribuyen según una Normal con media 0 y varianza constante.

[2] Por qué es el más adecuado está descrito en : N. R. Draper y H. Smith: *Applied Regression Analysis.* J. Wiley & Sons, 1998. Capítulo 5.

De hecho, gran parte de la estadística está estrechamente relacionada con el análisis de los cuadrados de los residuos (el análisis de la varianza, por ejemplo). Sin embargo, tampoco es prudente despreciar otros criterios, como el que hemos comentado de minimizar la suma de su valor absoluto. Por ejemplo, supongamos que tenemos los siguientes datos, en los que se ha cometido un error al introducir uno de los valores de Y

X	Y	
3	6	
6	5	
9	4	
12	13	← Error, debería ser 3
15	2	

El modelo que se obtiene ajustando por el método de los mínimos cuadrados es $Y = 6$ (Y=6+0X) porque el valor anómalo tiene una gran influencia sobre la ecuación obtenida. Sin embargo, ajustándolo con el criterio de minimizar la suma de los residuos en valor absoluto se obtiene $Y = 7 - (1/3)X$, que es el modelo que se obtendría si el valor erróneo se hubiera entrado correctamente

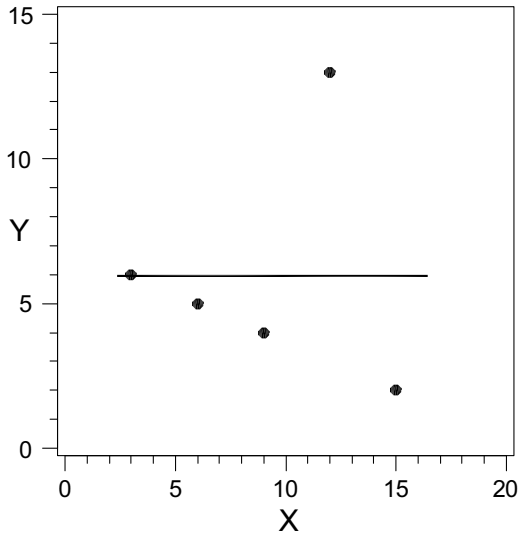


Figura 35.3. *Ajuste con el criterio de minimizar la suma de los cuadrados de los residuos*

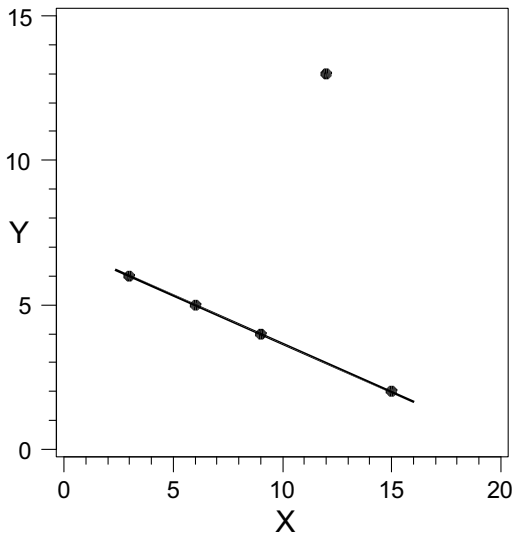


Figura 35.4. *Ajuste con el criterio de minimizar la suma de los residuos en valor absoluto*

Cuando se tiene una sola variable regresora es fácil identificar estos puntos anómalos en el análisis exploratorio inicial de los datos o en el posterior análisis de los residuos. Pero cuando se tienen muchas variables puede no ser tan fácil, y criterios de ajuste como este, robustos ante la presencia de valores anómalos, pueden ser contemplados en algunos casos, aunque solo sea como punto de vista complementario, especialmente ahora que los ordenadores están reduciendo cada vez más las dificultades de cálculo.

En general, las ventajas y desventajas de los dos criterios de ajuste son análogas a las que se comentaron en la pregunta 1 al comparar la media y la mediana.

36 Si los coeficientes de una ecuación de regresión son unos números concretos, ¿por qué se dice que son variables aleatorias?

Si usted coloca un conjunto de puntos en un plano, por ejemplo (2, 3), (4, 3), (6, 7) y (8, 7), de forma que aquí no hay variables aleatorias ni se considera ningún tipo de variabilidad, y desea calcular la ecuación de la recta que mejor se ajusta a esos puntos, puede hacerlo perfectamente y obtendrá: $Y = 1 + 0{,}8X$. Si los datos de partida son considerados como números fijos, los coeficientes de la ecuación también lo son, y asunto resuelto. Pero este no es el escenario que nos planteamos cuando abordamos este tipo de problemas desde un punto de vista estadístico.

Lo veremos con un ejemplo. Entre la altura y el peso de las personas hay una cierta relación, las personas más altas pesan más, en términos generales, que las más bajas, así que podríamos plantearnos hallar la ecuación que define esa tendencia al aumento de peso con el aumento de altura. Pero naturalmente no podremos medir y pesar a todas las personas del mundo, sino que tomaremos una muestra, por ejemplo de 20 individuos, a los que mediremos su peso y altura, y a partir de estos valores calcularemos la ecuación que andamos buscando[1].

Estaremos de acuerdo en que si en vez de las personas que hemos elegido hubiésemos tomado otras 20, la ecuación obtenida sería distinta. Esto es así porque dada una altura, el peso no es fijo, sino una variable aleatoria, que suponemos con distribución Normal. La Figura 36.1 muestra 6 situaciones de este tipo, en las que las alturas son valores que se han generado aleatoriamente de una N(170; 8) y los pesos se han calculado de la forma Peso = Altura – 100 + ε, siendo ε un número aleatorio tomado de una N(0; 5). Se ve claramente que aunque la relación es la misma (la media del peso es: altura – 100) la variabilidad en el peso para una altura dada, provoca que la recta no siempre sea la misma.

Que la recta no siempre sea la misma (ni en pendiente, ni en el punto de corte con el eje de la Y) es lo mismo que decir que sus coeficientes son variables aleatorias. Puede demostrarse que si la ε a que nos hemos referido anteriormente tiene unas ciertas propiedades[2], los coeficientes de la ecuación también siguen una distribución Normal, con una media que coincide con el valor que se obtendría si se utilizaran los datos de toda la población (bonita propiedad) y una desviación tipo que se puede calcular con base en la variación de los errores.

[1] Suponemos también que la relación entre el peso medio y altura es bien descrita por una línea recta.

[2] Básicamente que su media es cero y su varianza constante, así como que los errores asociados con cada punto son independientes entre sí.

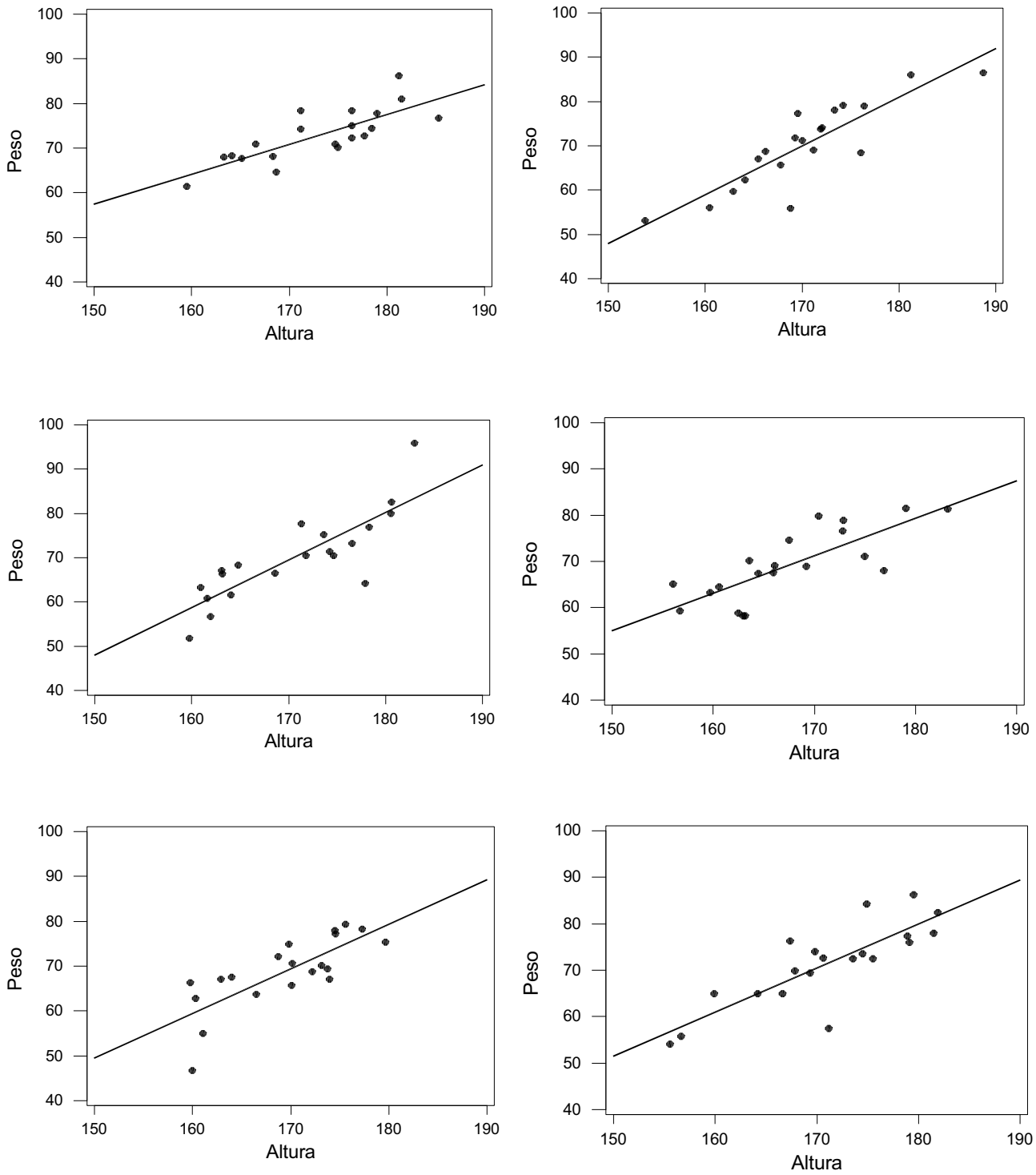


Figura 36.1. *Rectas que relacionan el peso y la altura de las personas, obtenidas a partir de 6 muestras distintas*[3]

Esto nos permite plantear pruebas de significación para los coeficientes (por ejemplo: ¿es la pendiente lo suficientemente distinta de cero para poder estar seguros de que realmente existe relación entre X e Y?). También se pueden calcular intervalos de confianza para la media de Y dado un valor de X (en nuestro caso la media del peso dado un valor de la altura). El valor medio del peso para una determinada altura es el que cae en la recta, pero como esta recta no es única, lo más adecuado es dar un intervalo, que estará relacionado con la superposición de todas las posibles rectas que se

[3] Los datos son ficticios. Las alturas son números aleatorios de una N(170; 8) y para los pesos se ha hecho: Peso = Altura – 100 + ε, siendo ε un valor de una N(0; 5)

podrían calcular. Los libros explican que este intervalo es más estrecho en el centro, en torno al punto $(\overline{X},\overline{Y})$ y que se va ensanchando hacia los extremos. Esto se puede comprobar superponiendo los 6 gráficos de la Figura 36.1, obteniéndose el resultado que puede observarse en la Figura 36.2. Para verlo todavía más claro lo hemos hecho también superponiendo 50 situaciones similares a las descritas, obteniéndose la Figura 36.3.

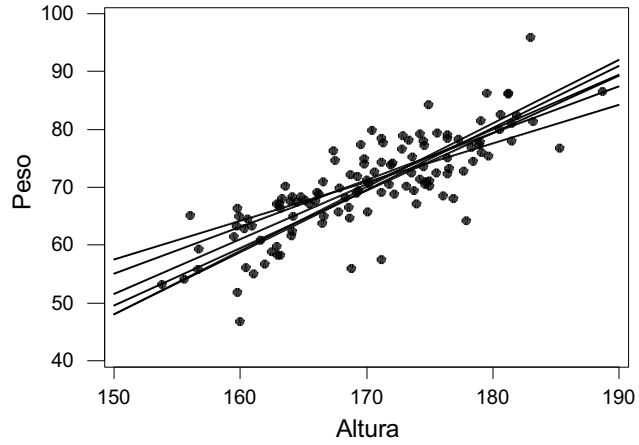


Figura 36.2. *Superposición de los 6 casos que aparecen en la Figura 36.1*

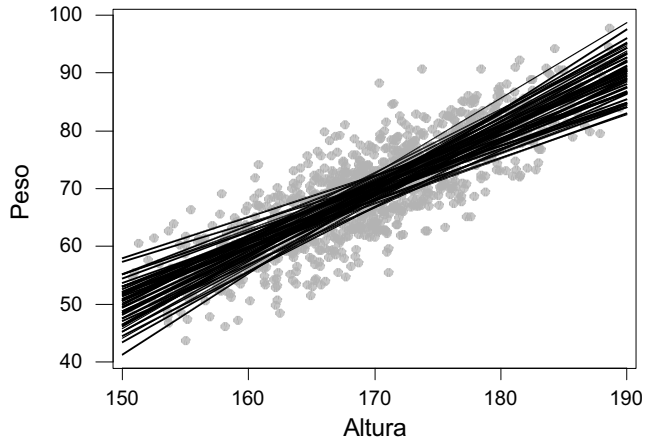


Figura 36.3. *Gráfico obtenido al superponer 50 gráficos del tipo de los que se incluyen en la Figura 36.1*

En definitiva, si ajustamos una recta para explicar la relación entre 2 variables, X e Y, a partir de una muestra de pares de valores de estas variables, la recta obtenida no se puede considerar fija y única para explicar la relación entre X e Y, ya que si la muestra hubiera sido otra, la recta también sería otra. La buena noticia es que haciendo determinados supuestos sobre el comportamiento de X e Y se pueden deducir cuáles son las distribuciones teóricas a que pertenecen los coeficientes de la recta, y esto nos permite realizar pruebas de significación o calcular intervalos de confianza para los valores que tomarían si se hubiera utilizado toda la población para ajustar la recta.

37 ¿Por qué cuando se ajusta una recta que pasa por el origen no se utiliza el coeficiente de determinación R^2 como medida de calidad del ajuste?

En una ecuación de regresión simple el coeficiente de determinación R^2 es la medida más utilizada de lo que se ha dado en llamar "bondad del ajuste". Primero vamos a repasar su significado en general para después entrar en el caso concreto de la recta por el origen.

Supongamos que X e Y son 2 variables relacionadas[1] de las que hemos obtenido los valores de la Figura 37.1.

X	Y
4	3
6	8
8	12
10	10
12	12
$\overline{X} = 8$	$\overline{Y} = 9$

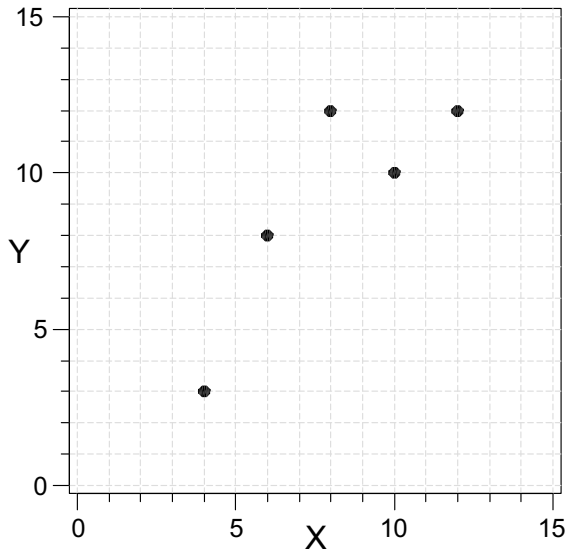


Figura 37.1. *Representación de los puntos en que se basa el ejemplo*

Si solo tuviéramos los valores de Y, sin conocer los correspondientes de X, la mejor predicción para Y, dado un valor de X, sería siempre la media de los valores conocidos de Y, ya que en este caso no podríamos obtener una ecuación de tipo $Y = f(X)$ que nos ayudara a mejorar las predicciones. Esta situación puede representarse a través de la ecuación de la recta: $\hat{Y} = b_0 \, (= \overline{Y})$.

Una forma de medir la "distancia" de la recta a los puntos es mediante la suma de los cuadrados de los residuos. De hecho, esta es la medida que utilizamos para seleccionar el mejor ajuste, que es el que la minimiza. ¿Cuánto vale en este caso en que no se han aprovechado para nada los valores de X? Con ayuda de Figura 37.2 es fácil determinar que la suma de los cuadrados de los residuos es igual a 56.

[1] Para ser rigurosos hay que decir que suponemos que la relación es del tipo $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, con ε_i independientes y distribuidos según una N(0; σ).

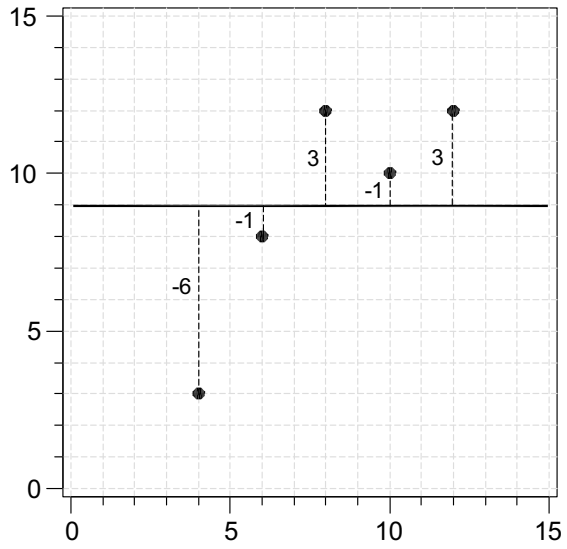


Figura 37.2. *Recta* $\hat{Y} = b_0$ *y valores de los residuos*

Veamos ahora lo que ocurre cuando utilizamos toda la información disponible para ajustar una recta por el método de los mínimos cuadrados. La ecuación obtenida es: $\hat{Y} = 1 + X$ (resultados tan redondos sólo salen cuando los números están preparados) y está representada en Figura 37.3 junto con los nuevos valores de los residuos. Puede observarse fácilmente que en este caso la suma de sus cuadrados vale 16.

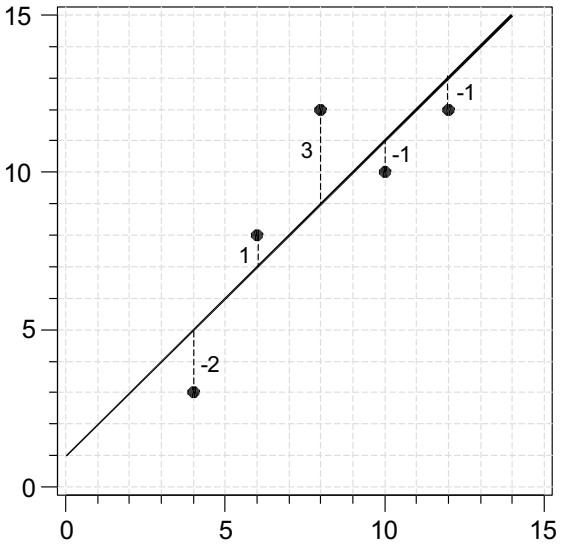


Figura 37.3. *Recta ajustada por mínimos cuadrados con los nuevos valores de los residuos*

La recta inicial, la que solo utiliza los valores de *Y*, da un valor de la suma de los cuadrados de los residuos al que llamaremos SCR_T (**S**uma de los **C**uadrados de los **R**esiduos **T**otal). Si gracias a *X* pudiéramos predecir perfectamente el comportamiento de *Y* (todos los puntos estuvieran sobre la recta) los residuos pasarían a ser cero y, por tanto, la suma de sus cuadrados también. No debemos aspirar a tanto cuando trabajemos con datos reales, pero sí es verdad que cuanto más sirva *X* para explicar el comportamiento de *Y*, más disminuirá la suma de los cuadrados de los residuos. A esta disminución la llamaremos SCR_E (**S**uma de los **C**uadrados de los **R**esiduos **E**xplicada)

El coeficiente de determinación R^2 se calcula de la forma:

$$R^2 = \frac{SCR_E}{SCR_T} = \frac{40}{56} = 0,714$$

La Figura 37.4 resume esta situación con los valores obtenidos en nuestro ejemplo.

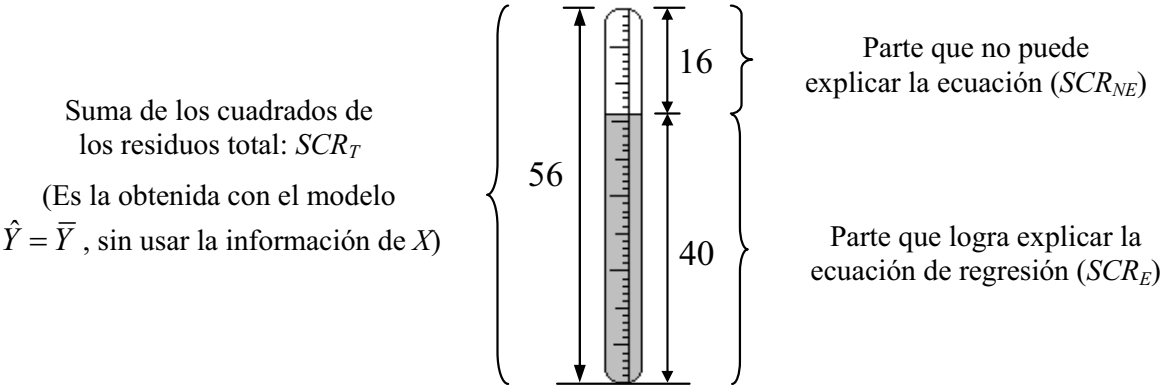


Figura 37.4. *"Termómetro" de medida del ajuste. R^2 = 40/56 = 0,714*

El valor de R^2 sirve para comparar distintos tipo de ajuste, especialmente en el caso de tener una sola variable regresora aunque en general puede usarse para comparar modelos con el mismo número de parámetros. También resulta que para el caso de ajuste a una línea recta, R^2 es el cuadrado del coeficiente de correlación entre X e Y,lo cual no deja de ser una curiosa coincidencia[2].

Centrémonos ahora en la recta que pasa por el origen. En primer lugar hay que aclarar (aunque seguramente no hace falta) que nos referimos a una recta que "se fuerza" a pasar por el origen. Si pasa de una forma natural no hay nada nuevo y vale todo lo dicho para el caso general.

Con los datos de nuestro ejemplo, si se fuerza a la recta a pasar por el origen, su ecuación queda: $\hat{Y} = \frac{11}{10} X$. El paquete de software estadístico Minitab no da el valor de R^2, pero evidentemente se puede calcular de la forma descrita (aunque tampoco hace falta hacerlo a mano porque Excel sí lo da), obteniéndose R^2 = 0,70. ¿Informa este valor sobre la calidad del ajuste? Sí, pero…

- No permite comparar con otras rectas. La recta "libre" ajustada por mínimos cuadrados tiene un R^2 del 71,4%. Si la forzamos a pasar por el origen es porque consideramos que así refleja mejor la realidad y es, por tanto, un modelo mejor, aunque su R^2 sea menor.

[2] En la teoría estadística también se acaba descubriendo con sorpresa que las medidas y variables aleatorias más influyentes están emparentadas entre sí.

- Al ser un modelo no ajustado por el método de los mínimos cuadrados, no cumple con las propiedades que se describen para R^2. En concreto, en este caso R^2 no es el cuadrado del coeficiente de correlación lineal entre X e Y,

- El referente para juzgar su tamaño ya no es el intervalo (0; 1), pues el valor $R^2 = 1$ solo se lograría cuando pasa por el origen sin forzarla y todos los puntos están alineados.

Puede prestarse a confusión llamar R^2 a algo que, aunque calculado de la misma forma, no tiene sus características ni sus propiedades. Por eso Minitab no lo da, aunque otros como Excel sí.

38 ¿Por qué cuando se comparan ecuaciones de regresión con distinto número de variables regresoras no se utiliza R^2 sino el llamado R^2 ajustado?

Supongamos que disponemos de la siguiente muestra de datos (los mismos que usamos para responder a la pregunta anterior) sobre las características X e Y de individuos de cierta población, y que deseamos construir un modelo para predecir la media de *Y* cuando la característica *X* toma algún valor específico *x*.

X	Y
4	3
6	8
8	12
10	10
12	12
$\overline{X}=8$	$\overline{Y}=9$

Vamos a comparar el ajuste obtenido utilizando 5 modelos. El primer modelo no incluye la variable *X*, es decir, que la predicción para *Y* es una constante que no depende de *X*. El segundo se construirá escogiendo la mejor[1] recta entre todas las posibles rectas, el tercero escogiendo la mejor parábola entre todas las parábolas y para el cuarto y quinto modelo elegiremos los mejores polinomios de tercer y cuarto grado.

Modelo 1: $y_i = \mu + \varepsilon_i$

Para este modelo la estimación de *Y* corresponde a la media aritmética de los datos, pues es justamente la media la que minimiza la suma de los cuadrados de los residuos.

Predicción de la media de *Y* para cualquier valor de *X*: $\hat{y} = \overline{y} = 9$

Suma de los cuadrados de los residuos con el modelo usado: $\sum e_i^2 = 56$

Suma de los cuadrados de los residuos sin hacer uso de *X*: $\sum (y_i - \overline{y})^2 = 56$

Coeficiente de determinación:[2]

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \overline{y})^2} = 0$$

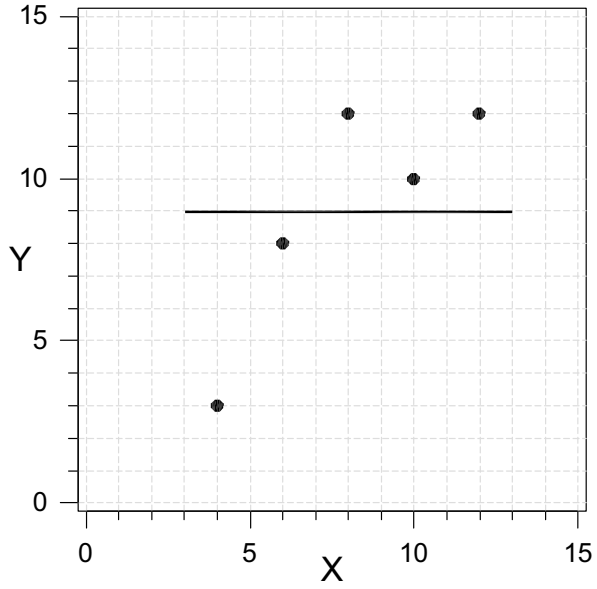


Figura 38.1. *Ajuste a una constante*

[1] Cuando hablamos de "mejor", queremos decir "el que minimiza la suma de los cuadrados de los residuos".

[2] La expresión del coeficiente de determinación se utiliza escrita de varias formas. Para contestar la pregunta anterior utilizamos $R^2 = \frac{SCR_E}{SCR_T}$, siendo SCR_E la suma de los cuadrados de los residuos

Modelo 2: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Predicción de la media de *Y* cuando $X = x_i$: $\hat{y} = 1 + x_i$

Suma de los cuadrados de los residuos con el modelo usado: $\sum e_i^2 = 16$

Suma de los cuadrados de los residuos sin hacer uso de *X*: $\sum (y_i - \overline{y})^2 = 56$

Coeficiente de determinación:

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \overline{y})^2} = 1 - \frac{16}{56} = 0,714$$

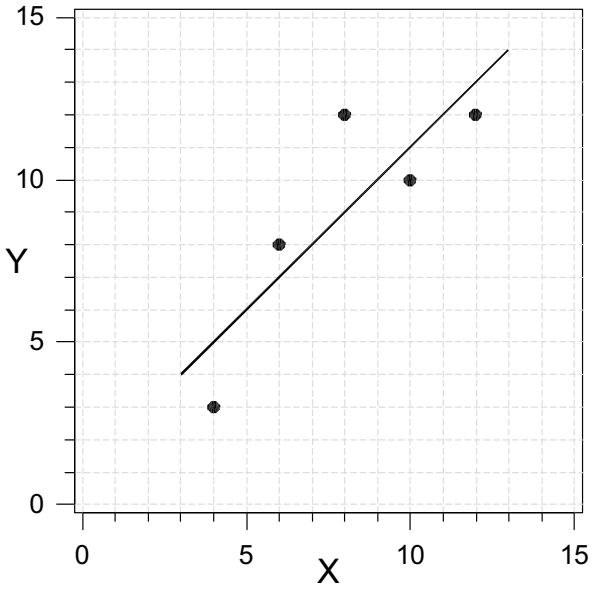


Figura 38.2. *Ajuste a una recta*

Modelo 3: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$

$\hat{y} = -11 + 4,429 \cdot x_i - 0,214 \cdot x_i^2$

$\sum e_i^2 = 5,7$

$\sum (y_i - \overline{y})^2 = 56$

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \overline{y})^2} = 1 - \frac{5,7}{56} = 0,898$$

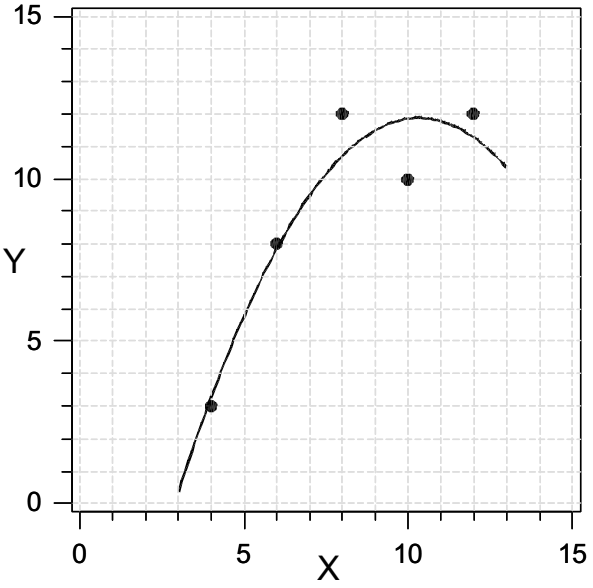


Figura 38.3. *Ajuste a un polinomio de 2º grado*

A priori ya sabíamos que la mejor ecuación con este modelo tendría un R^2 mayor o igual que el de la ecuación obtenida usando el modelo 2, puesto que el valor de b_2 que minimiza la suma de los cuadrados de los residuos es $b_2 = -0,214$, y el modelo 2 es un caso particular del modelo 3 en el que $b_2 = 0$. Si el valor de b_2 que minimiza la suma de los cuadrados de los residuos es –0,214, cualquier otro (incluido el cero) dará una suma de cuadrados mayor.

---

explicada por la regresión y SCR_T a la suma de los cuadrados de los residuos total. Con la notación que ahora estamos usando, $SCR_E = \sum (y_i - \overline{y})^2 - \sum e_i^2$ y $SCR_T = \sum (y_i - \overline{y})^2$, de donde se deduce que

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \overline{y})^2}$$

Algo similar ocurrirá cuando ajustemos el modelo correspondiente a polinomios de grado superior. Sabemos que la suma de los cuadrados del error para el mejor modelo basado en un polinomio de tercer grado será menor o igual que 5,7, puesto que todos los polinomios de segundo grado están contenidos en los de tercero haciendo $\beta_3 = 0$.

Modelo 4:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$$

$$\hat{y} = -32 + 13,7 \cdot x_i - 1,46 \cdot x_i^2 + 0,0521 \cdot x_i^3$$

$$\sum e_i^2 = 3,2$$

$$\sum (y_i - \bar{y})^2 = 56$$

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{3,2}{56} = 0,946$$

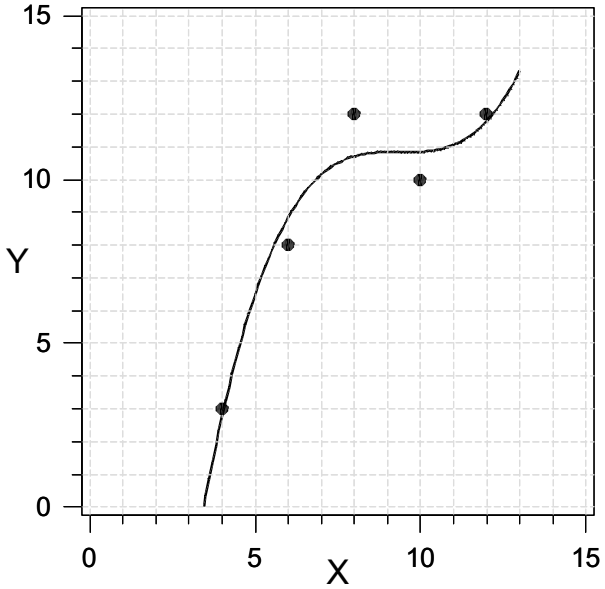


Figura 38.4. *Ajuste a un polinomio de 3r grado*

Para este modelo la suma de cuadrados de los residuos resultó ser 3,2, es decir, que con referencia al modelo que no incluye la variable predictora, bajó de 56 a 3,2, Esto significa que disminuyó en 52,8 unidades, lo cual representa el 94,6%

Por último veamos el ilustrativo caso del polinomio de cuarto grado, en el que sin hacer ningún cálculo podemos anticipar que tendrá una suma de cuadrados de residuos nula. Es decir, sabemos a priori que todos los puntos caerán sobre la curva encontrada y que por lo tanto su coeficiente de determinación será del 100%, siendo su ajuste "perfecto".

Modelo 5:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 \varepsilon_i$$

$$\hat{y} = 85 - 55,21 x_i - 12,84 x_i^2 - 1,20 x_i^3 + 0,03 x_i^4$$

$$\sum e_i^2 = 0$$

$$\sum (y_i - \bar{y})^2 = 56$$

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} = 1 - 0 = 1$$

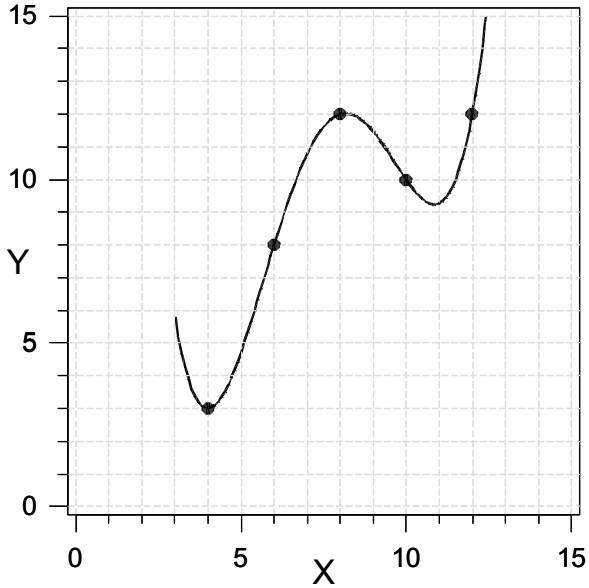


Figura 38.5. *Ajuste a un polinomio de 4º grado*

En este último modelo ocurre una gran paradoja. Sin necesidad de conocer los datos, ya se sabía de antemano que el ajuste sería perfecto, pues así como por dos puntos pasa una única recta, por tres puntos pasa una parábola de segundo grado, por cuatro puntos siempre puede pasar una parábola cúbica, y por nuestros cinco puntos pasa un polinomio de grado cuatro. Siempre por n+1 puntos pasará un polinomio de grado n, y el ajuste será "perfecto" sin importar cuáles sean los datos. Esto es un verdadero adefesio, ya que este tipo de ecuaciones no tienen ningún poder de predicción, que es lo realmente importante[3].

Para evitar ese problema se ha corregido la definición de R^2, dando origen al coeficiente de determinación ajustado (o corregido) R^2_{Aj}, dividiendo el numerador y el denominador de la expresión de R^2, por sus correspondientes grados de libertad, así:

$$R^2_{Aj} = 1 - \frac{\dfrac{\sum e_i^2}{n-p}}{\dfrac{\sum (y_i - \bar{y})^2}{n-1}}$$

← Varianza de los residuos

← Varianza muestral de y

Haciendo las sustituciones correspondientes podemos relacionar los 2 coeficientes de la forma:

$$R^2_{Aj} = 1 - \frac{n-1}{n-p}\left(1 - R^2\right)$$

Podemos ahora calcular el coeficiente de determinación ajustado para cada uno de los modelos que hemos considerado:

Familia de modelos	R^2	R^2_{aj}
$y_i = \mu + \varepsilon_i$	0%	0%
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$	71,4%	61,8%
$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$	89,8%	79,6%
$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$	94,6%	78,4%
$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 \varepsilon_i$	100%	Indeterminado

Nótese cómo el coeficiente de determinación ajustado baja su valor cuando al modelo de grado 2 se le adiciona una componente de grado 3. Esto indica que la ganancia

[3] Usted puede construir un modelo que explique perfectamente la cotización en bolsa de las 5 grandes empresas que desee. Tome las 5 cotizaciones de hoy (que serán las Y) seleccione ahora las temperaturas máximas de ayer en 5 capitales eurpeas (X_1), el tipo de cambio de las 5 monedas que usted elija frente al dólar (X_2), los precios de 5 materias primas (las que quiera) en el mercado de Londres (X_3) y la temperatura media en Barcelona en los últimos 5 días (X_4). Obtenga el modelo que explica Y en función de X_1, X_2, X_3, X_4. ¿Ajuste perfecto! Lástima que para hoy el modelo llega tarde, y para mañana es totalmente inútil.

adicional en la disminución de la suma de cuadrados de los residuos no se compensa con la pérdida de un grado de libertad para el error. El efecto sería el mismo si en lugar de términos polinómicos planteamos el modelo con una, dos, tres, cuatro y cinco variables distintas.

Cuando la relación *n*/*p* se hace grande, es decir, que el número de datos supera mucho al número de parámetros, los dos coeficientes se acercan en sus valores. Veamos las siguientes situaciones.

Si con 10 datos se construyera un polinomio de grado 8 y resultara un coeficiente de determinación R^2= 90%, podrá dar la falsa impresión de que estamos ante un buen modelo. Sin embargo, utilizando la fórmula vista anteriormente que relaciona los 2 coeficientes, tenemos:

$$R^2_{Aj} = 1 - \frac{10-1}{10-9}\left(1-0,9\right) = 0,1$$

Es decir, nos indica que en esas condiciones el valor creíble del coeficiente de determinación es el 10%. El ajuste es bastante pobre.

Supongamos ahora la misma situación anterior donde lo único distinto es que todos los cálculos y estimaciones se realizaron con *n*=90 datos.

$$R^2_{Aj} = 1 - \frac{90-1}{90-9}\left(1-0,9\right) = 0,89$$

Pasamos del 90 al 89%, es decir, que tuvo un cambio casi despreciable. Note que en esta ocasión se cumple la recomendación empírica de que hayan 10 datos por cada parámetro, es decir, que la razón *n*/*p* valga por lo menos 10.

39 ¿Cómo se pueden utilizar e interpretar variables cualitativas en una ecuación de regresión?

Resulta claro que las variables cualitativas no se pueden introducir tal cual en una ecuación de regresión. Por ejemplo, si el día de la semana es una posible variable explicativa y tenemos los días codificados como Lunes = 1, Martes = 2, ..., es evidente que esta variable no se puede considerar así, ya que el modelo entendería que el domingo es 7 veces el lunes, cosa que evidentemente es falsa. Sin embargo, sí hay forma de poder utilizar variables cualitativas, lo veremos a través de un ejemplo sencillo.

En la Tabla 39.1 tenemos los datos correspondientes al peso (en kg), la altura (en m) y el sexo de 20 individuos (los datos son ficticios, hemos puesto unos valores que ilustren de forma clara la situación que queremos mostrar). Se trata de encontrar una ecuación para explicar el peso medio en función de la altura y el sexo. También interesa sacar conclusiones sobre cómo el sexo afecta a la forma de la relación entre peso y altura.

Tabla 39.1. *Datos del ejemplo*

Individuo	Peso (kg)	Altura (m)	Sexo Hom/Mu
1	55,0	1,60	M
2	91,7	1,79	H
3	69,7	1,67	H
4	56,8	1,64	M
5	66,3	1,66	H
6	59,5	1,70	M
7	71,9	1,69	H
8	97,6	1,83	H
9	58,0	1,63	M
10	54,3	1,59	M
11	82,0	1,74	H
12	75,9	1,68	H
13	64,0	1,70	M
14	73,0	1,78	M
15	70,0	1,74	M
16	78,3	1,70	H
17	84,2	1,77	H
18	98,6	1,83	H
19	60,5	1,66	M
20	70,9	1,66	H

Lo primero será echar un vistazo a los datos a través de un gráfico. En este caso basta con un diagrama bivariante usando distinto símbolo según el sexo. Utilizando Minitab se obtiene el gráfico de la Figura 39.1 en el que se aprecia claramente que hombres y mujeres están agrupados en torno a rectas distintas (insistimos en que los datos son ficticios).

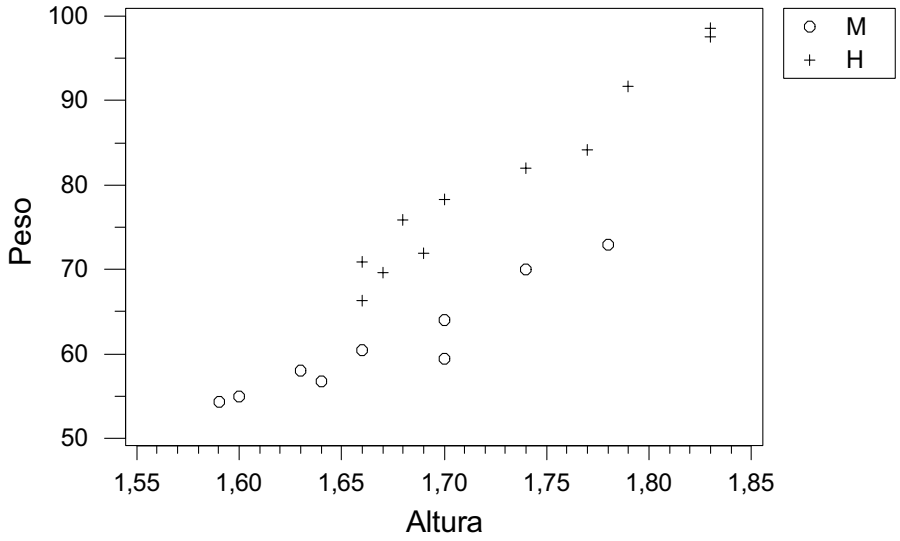


Figura 39.1. *Diagrama bivariante de los datos del ejemplo estratificados según el sexo*

Para determinar la ecuación primero daremos valores numéricos a la variable 'sexo'. Lo más cómodo es usar 0 y 1. En este caso convenimos: Mujer = 0 y Hombre = 1. Siguiendo con Minitab, obtenemos:

```
The regression equation is
Peso = - 169 + 138 Altura + 11,4 Sexo

Predictor        Coef     SE Coef          T        P
Constant      -168,72       19,09      -8,84    0,000
Altura         137,60       11,40      12,07    0,000
Sexo           11,434       1,563       7,31    0,000

S = 3,152       R-Sq = 95,1%     R-Sq(adj) = 94,6%
```

Figura 39.2. *Salida de Minitab (parcial) cuando se ajusta un modelo del peso en función de altura y sexo*

Del resultado anterior se puede deducir que sí hay diferencias de modelo debido al sexo, ya que el coeficiente de esta variable es significativo. Sustituyéndola por sus valores 0 y 1, se obtienen los modelos específicos para mujeres y hombres:

Modelo para mujeres (sexo = 0): `Peso = -169 + 138 Altura`
Modelo para hombres (sexo = 1): `Peso = -157,3 + 138 Altura`

Así de sencillo. Aunque ajustando el modelo de esta forma la pendiente de la recta es la misma para los 2 valores de la variable cualitativa, ya que solo permite valorar la diferencia de b_0 (este es el valor que cambia al cambiar el valor de la variable cualitativa). Colocando las rectas obtenidas sobre el diagrama bivariante se tiene el resultado que se representa en la Figura 39.3.

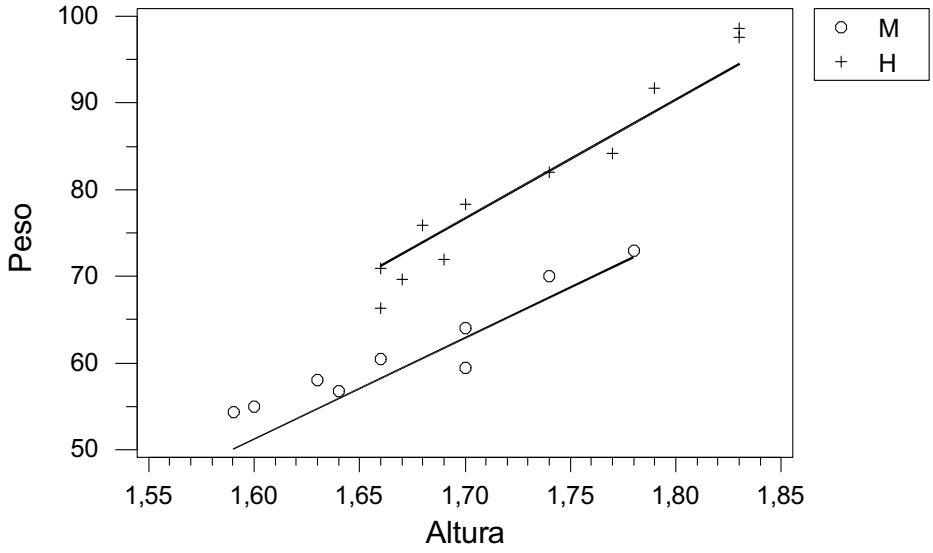


Figura 39.3. *Rectas ajustadas para hombres y mujeres con la misma pendiente*

Si consideramos el caso más general en el cual la pendiente de las rectas puede ser distinta, deberemos incluir en el modelo el producto de la variable binaria por las otras variables regresoras (en este caso solo la altura). De esta forma se obtiene el nuevo modelo que se indica en la Figura 39.4.

```
The regression equation is
Peso = - 102 + 97,9 Altura - 106 Sexo + 69,4 Altura*Sexo

Predictor        Coef     SE Coef          T        P
Constant      -102,34       20,51      -4,99    0,000
Altura          97,88       12,27       7,98    0,000
Sexo          -106,33       27,53      -3,86    0,001
Altura*S        69,44       16,22       4,28    0,001

S = 2,218       R-Sq = 97,7%     R-Sq(adj) = 97,3%
```

Figura 39.4. *Salida de Minitab (parcial) cuando se añade el término de interacción altura × sexo*

La significación del término Altura×Sexo pone de manifiesto la diferencia en las pendientes de las rectas. Los modelos que se obtienen en este caso son:

Modelo para mujeres (sexo = 0): `Peso = -102 + 97,9 Altura`
Modelo para hombres (sexo = 1): `Peso = -208 + 167,3 Altura`

Su representación gráfica puede verse en la Figura 39.5. Seguro que ahora nos quedamos mucho más tranquilos con los ajustes obtenidos.

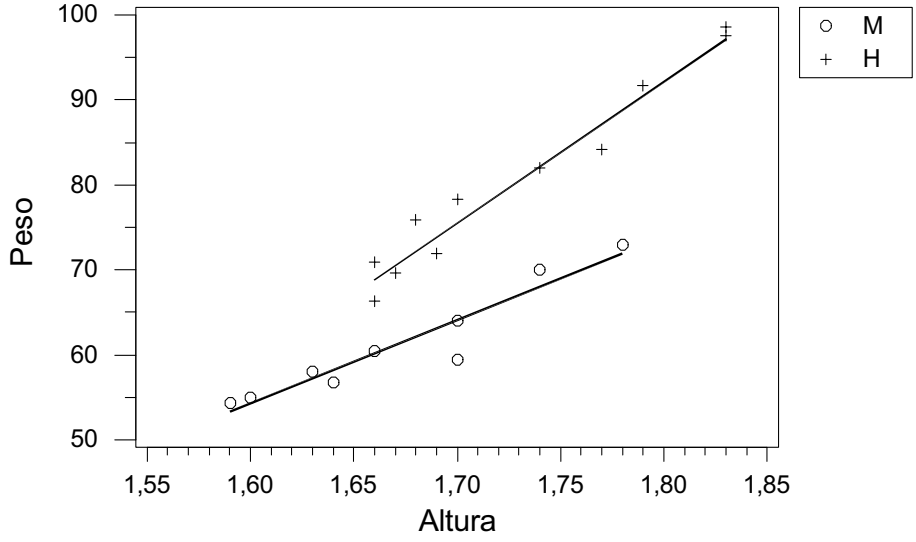


Figura 39.5. *Rectas correspondientes a los modelos que incluyen el término de interacción sexo × altura*

¿Y si la variable cualitativa tiene más de 2 valores distintos? Veamos cómo se podría introducir la variable "día de la semana" suponiendo que no interacciona con las cuantitativas, que supondremos que son dos (a título de ejemplo).

Variables regresoras: cuantitativas		Variables regresoras: cualitativa	Variables auxiliares						Respuesta
X_1	X_2	**Día**	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Y
X_{11}	X_{21}	L	1	0	0	0	0	0	Y_1
:	:	:	:	:	:	:	:	:	:
:	:	L	1	0	0	0	0	0	:
:	:	M	0	1	0	0	0	0	:
:	:	:	:	:	:	:	:	:	:
:	:	M	0	1	0	0	0	0	:
:	:	Mi	0	0	1	0	0	0	:
:	:	:	:	:	:	:	:	:	:
:	:	Mi	0	0	1	0	0	0	:
:	:	J	0	0	0	0	0	0	:
:	:	:	:	:	:	:	:	:	:
:	:	J	0	0	0	1	0	0	:
:	:	V	0	0	0	0	1	0	:
:	:	:	:	:	:	:	:	:	:
:	:	V	0	0	0	0	1	0	:
:	:	S	0	0	0	0	0	1	:
:	:	:	:	:	:	:	:	:	:
:	:	S	0	0	0	0	0	1	:
:	:	D	0	0	0	0	0	0	:
:	:	:	:	:	:	:	:	:	:
X_{1n}	X_{2n}	D	0	0	0	0	0	0	Y_n

Tabla 39.2. *Uso de variables auxiliares para poder considerar el día de la semana como variable regresora*

La regla general es que una variable cualitativa con k valores distintos debe sustituirse por k-1 variables binarias que toman valores 0 y 1. En el caso de la variable sexo del primer ejemplo, como presenta 2 valores distintos (hombre y mujer) la hemos sustituido por otra con valores 0 y 1. En el caso de la variable "día de la semana", la sustituiremos por 6 variables binarias, tal como se indica en la Tabla 39.2.

La ecuación obtenida con las 2 variables cuantitativas y las 6 auxiliares será:

$$Y = b_0 + b_1X_1 + b_2X_2 + \lambda_1Z_1 + \lambda_2Z_2 + \ ... \ + \lambda_6Z_6$$

Y tendremos los modelos:

Para el lunes:	$Y = b_0 + \lambda_1 + b_1X_1 + b_2X_2$
Para el martes:	$Y = b_0 + \lambda_2 + b_1X_1 + b_2X_2$
:	
Para el domingo:	$Y = b_0 + b_1X_1 + b_2X_2$

Obsérvese que lo que tenemos son planos paralelos. Si deseamos que los coeficientes b_1 o b_2 puedan cambiar con el día habrá que añadir los productos tanto de X_1 como de X_2 con cada una de las variables auxiliares. (En total serían 18 nuevos coeficientes, los 6 que ya teníamos más 12 correspondientes a los productos).

Si esto resulta demasiado complicado, o si el número de observaciones de que se dispone no da para estimar con solvencia tantos coeficientes, una opción podría ser clasificar los días solo como laborables y festivos. De esta forma no solo el análisis es más sencillo, sino que la presentación de los resultados se puede realizar de forma más clara y compacta, lo cual también es importante.

40 ¿Por qué del conjunto de variables candidatas a entrar en un modelo de regresión no necesariamente se seleccionan las que están más correlacionadas con la variable dependiente *Y*?

Es muy fácil de entender con un ejemplo. Suponga que tiene que presentarse al examen de una asignatura cuyo programa consta de 100 temas y resulta que usted no sabe ninguno. Pero no todo está perdido, las reglas de este examen dicen que usted puede llevar compañeros de clase como asesores y, además, puestos a suponer, supondremos que usted sabe qué temas conoce cada uno de sus compañeros.

Si pudiera llevar a un solo asesor, ¿a cuál elegiría? La respuesta es fácil: evidentemente, al que más temas sepa. Supongamos que este es Pablo, que sabe 85 temas.

A la hora de elegir el segundo, una opción que parece razonable es elegir el segundo que más sabe, por ejemplo, Alberto, que sabe 75 temas. El problema de esta estrategia es que puede ocurrir que los 75 temas de Alberto ya estén incluidos entre los 85 que sabe Pablo y, por tanto, que Alberto no aporte absolutamente nada cuando ya se cuenta con Pablo.

Una estrategia mejor es seleccionar el segundo buscando el que más sabe, no de todo el temario, sino de lo que le falta por saber al primero. Seguro que el lector estará de acuerdo con nosotros. Incluso puede ocurrir que en la mejor pareja de asesores no esté el que más sabe, ya que pueden haber 2 que se complementen perfectamente, de modo que uno sepa 55 temas y el otro los 45 restantes, mientras que el que más sabe no se complemente bien con ningún otro.

El símil con la obtención del modelo de regresión está claro. La primera variable en ser elegida como regresora es la más correlacionada con *Y*, es decir, la que mejor explica su comportamiento. Pero la segunda no será la segunda más correlacionada con *Y*, ya que lo que explica esta quizá ya lo explicaba la primera. Una vez seleccionada una variable el criterio es: seleccionar la que mejor explique lo que falta por explicar.

También hará falta que cumpla con otros requisitos, como que una vez en el modelo su coeficiente supere un cierto nivel de significación. La selección y el descarte de variables siguiendo estos criterios lo resuelven muy bien los paquetes de software estadístico utilizando la estrategia de selección de modelos denominada “paso a paso” (“*stepwise*”).

Diseño de experimentos

41 ¿Por qué no es una buena estrategia ir moviendo las variables una a una cuando se trata de estudiar experimentalmente cómo estas afectan a una respuesta?

La verdad es que a primera vista esta estrategia parece razonable. Si se trata de optimizar un proceso sobre el que se piensa que pueden influir diez variables la estrategia sería fijar nueve de ellas e ir probando diferentes valores de la décima hasta encontrar aquel que maximiza la respuesta (suponiendo que este sea el objetivo). A continuación, fijar esta variable a su "mejor" valor y experimentar cambiando los valores de una de las nueve restantes. El procedimiento continuaría hasta haber experimentado con las diez variables. Aparentemente este método está bien organizado, conduce al óptimo y además tiene la gran ventaja de que los resultados son muy fáciles de analizar.

Pero no nos engañemos. Veamos gráficamente cómo funciona este procedimiento en un caso con solo dos variables (con diez resulta imposible visualizarlo, pero la situación es totalmente análoga). Se desea maximizar la cantidad de producto obtenido como resultado de una reacción química sobre la que se sabe que hay dos variables que pueden resultar influyentes: la temperatura del reactor (habitualmente fijada a 130ºC) y el tiempo de reacción (habitualmente 90 minutos); la cantidad que se obtiene en estas condiciones es de 61 gr.

El procedimiento consiste en mantener fija la temperatura en su valor habitual y probar distintos valores del tiempo, con lo que se obtiene una cantidad máxima de 65,5 gr. que corresponde a un tiempo de 100 minutos. Una cierta mejora. Siguiendo con el procedimiento se fija el tiempo a 100 minutos y se experimenta con diversos valores de la temperatura. La nueva cantidad máxima es de 82 gr. correspondiente a una temperatura de 143ºC (ver Figura 41.1).

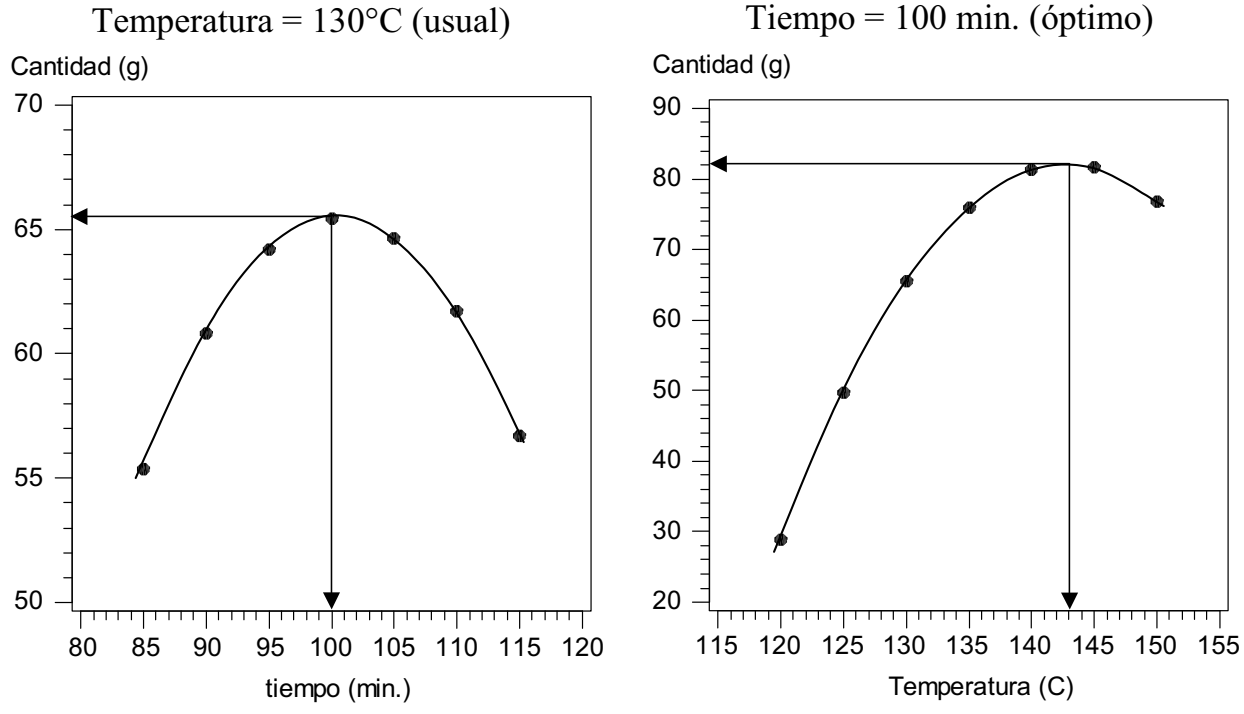


Figura 41.1. *Cantidad de producto obtenido moviendo una variable cada vez*

Así pues, una vez concluido el experimento, se ha conseguido aumentar la cantidad producida en 21 g. Pero, ¿hemos determinado realmente las condiciones óptimas de producción? En la Figura 41.2 la cantidad obtenida está representada por curvas de nivel en función del tiempo y de la temperatura, y se han representado también los puntos que corresponden a los experimentos anteriores. Es evidente que no se ha alcanzado el óptimo.

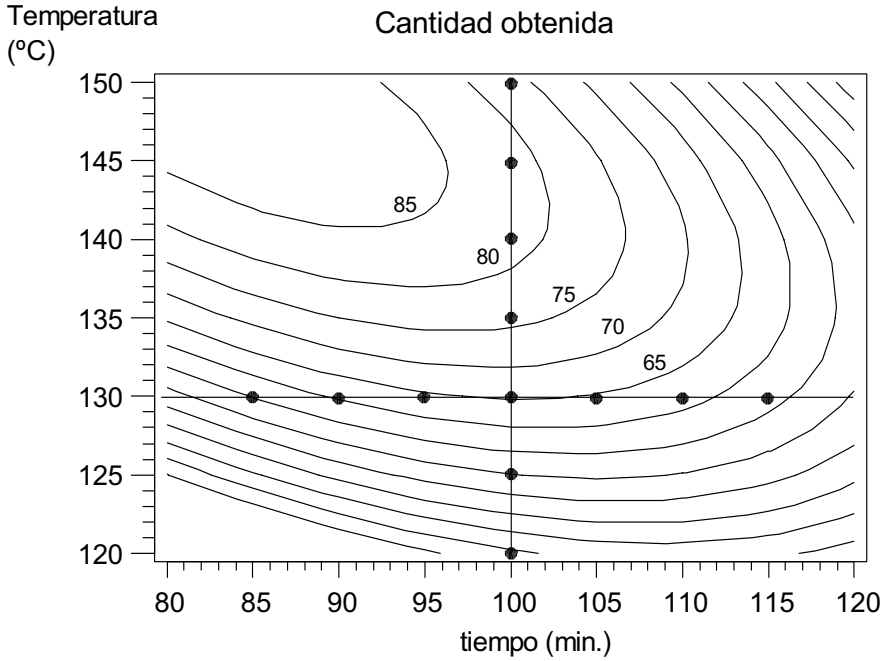


Figura 41.2. *Curvas de nivel que representan la cantidad de producto obtenido en función del tiempo y de la temperatura*

Este tipo de estrategia tiene la ventaja de que es fácil de entender y parece razonablemente buena (cuesta convencer a algunas personas de que hay alternativas mejores), pero tiene algunos inconvenientes que desaconsejan su uso:

1. Proporciona falsos óptimos, que pueden estar lejos del real dependiendo de dónde se empiece la experimentación.

2. Es una estrategia lenta de acercamiento al óptimo. Podríamos volver a determinar el tiempo que maximiza la respuesta para la nueva temperatura encontrada, y después hacer lo mismo con el tiempo, y así sucesivamente, trazando líneas sobre la superficie de respuesta, pero es evidente que este no es el camino más rápido de acercarse al óptimo.

3. No permite detectar interacciones entre los factores, y esta es una limitación muy importante para entender cómo las variables afectan a la respuesta.

Para estudiar a través de la experimentación cómo un conjunto de variables afectan a una respuesta, tenemos una alternativa mucho mejor. Son los diseños experimentales que conocemos con el nombre de diseños factoriales. A primera vista parecen un lío difícil de desenmarañar, pero cuando más se conocen más se aprecian. Si el tema le parece interesante y tiene la oportunidad, vale la pena dedicar un tiempo a conocerlos.

42 ¿Cómo es posible estudiar por separado el efecto de cada una de las variables que afectan a una respuesta si, tal y como se hace en los diseños factoriales, se mueven todas a la vez?

Los diseños factoriales presentan unas amplias posibilidades de uso en la optimización de productos y procesos industriales, siendo una de sus características más relevantes el que en cada tanda de experimentación (normalmente el plan completo está formado por varias tandas siguiendo una estrategia secuencial) se realizan experimentos en todas las combinaciones de valores de las variables con que se experimenta.

Por ejemplo, si tenemos 3 variables (a las que llamamos factores) y cada una de ellas va a ser estudiada en 2 valores distintos[1], que podemos representar codificados con los signos – y +, en total habrá que realizar experimentos en 8 condiciones distintas, tal y como se indica en la Figura 42.1[2]. La expresión del número de experimentos que hay que hacer (número de niveles elevado a número de factores) da nombre al diseño y, por tanto, este será un diseño 2^3.

Valores de los factores			Combinación resultante A	B	C	Respuestas
C (–)	B (–)	A (–)	–	–	–	y_1
		A (+)	+	–	–	y_2
	B (+)	A (–)	–	+	–	y_3
		A (+)	+	+	–	y_4
C (+)	B (–)	A (–)	–	–	+	y_5
		A (+)	+	–	+	y_6
	B (+)	A (–)	–	+	+	y_7
		A (+)	+	+	+	y_8

Figura 42.1. *Combinaciones de valores de los factores que conforman las 8 condiciones de experimentación de un diseño factorial con 3 factores a 2 niveles (diseño 2^3)*

Yendo ya a la pregunta planteada, sí es verdad que si movemos todas las variables a la vez parece que va a ser imposible separar la influencia de cada una de ellas, pero sin embargo, esto es perfectamente posible. Para verlo lo mejor será seguir un ejemplo.

[1] En general, en cada tanda de experimentación, cada factor toma solo 2 valores distintos. Estos diseños, que denominamos "a 2 niveles", presentan una excelente relación entre información obtenida y esfuerzo experimental.

[2] Se ha empezado el árbol de generación de niveles a partir del último factor, en este caso el C, para que aparezca el llamado orden estándar de la matriz de diseño (relación de condiciones en las que se va a experimentar). No hace falta escribir el árbol para generar la matriz del diseño, basta con alternar los signos – y + en la columna correspondiente al primer factor, en la siguiente columna alternarlos agrupados de 2 en 2, en la siguiente de 4 en 4, a continuación de 8 en 8, etc.

Supongamos que en un proceso de fabricación de galletas se desea maximizar lo crujientes que resultan y se considera que en esta característica pueden influir 3 variables: la temperatura del horno, el tiempo de horneado y el tipo de mantequilla utilizada. Supongamos también que se ha decidido (basándose en criterios técnicos) experimentar con temperaturas de 200 y 220º C, tiempos de 20 y 25 minutos y 2 tipos de mantequilla que denominaremos A y B.

Siguiendo la pauta marcada en la Figura 42.1 podemos escribir las condiciones en las que deberemos realizar nuestros 8 experimentos. Estas condiciones están indicadas en la Tabla 42.1 junto con el resultado obtenido en cada una de ellas.

Tabla 42.1. *Condiciones de experimentación y resultados obtenidos*

nº	Temperatura	tiempo	Tipo de Mantequilla	Medida de crujiente
1	200	20	A	7,9
2	220	20	A	9,7
3	200	25	A	7,5
4	220	25	A	9,2
5	200	20	B	6,4
6	220	20	B	8,4
7	200	25	B	7,3
8	220	25	B	9,0

Si representamos estos resultados en los 8 vértices de un cubo de forma que las coordenadas de cada vértice indican las condiciones en que se ha obtenido el resultado (Figura 42.2), se observan unas peculiares características de equilibrio y simetría debido a una propiedad de estos diseños que conocemos con el nombre de ortogonalidad.

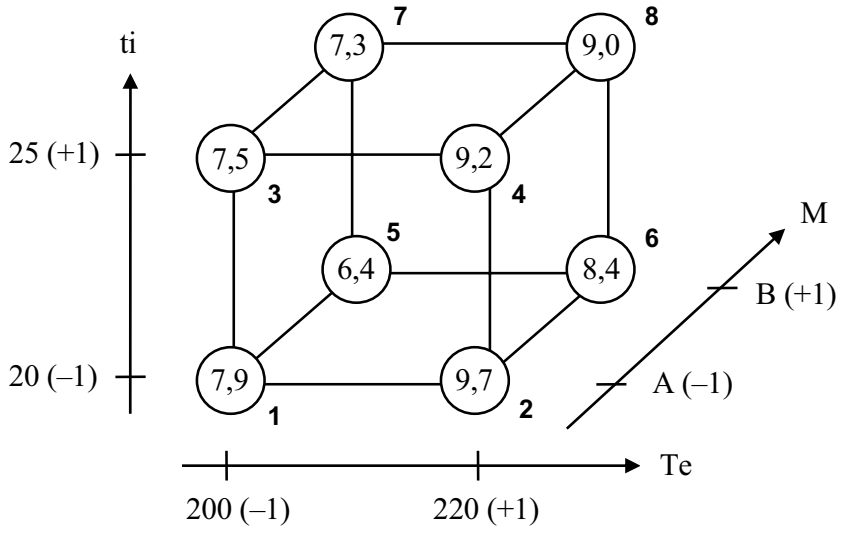


Figura 42.2. *Representación gráfica de variables y respuesta*

Podemos observar que entre las condiciones 2 y 1 la *única diferencia* en los valores de los factores es la temperatura, y lo mismo ocurre entre las condiciones 4 y 3, 6 y 5, y 8 y 7. El promedio de estas diferencias es el efecto principal de la temperatura, que podemos calcular de la forma:

$$\text{Te} = \frac{(y_2 - y_1) + (y_4 - y_3) + (y_6 - y_5) + (y_8 - y_7)}{4} =$$

$$= \frac{y_2 + y_4 + y_6 + y_8}{4} - \frac{y_1 + y_3 + y_5 + y_7}{4} = 1{,}8$$

Es decir, al cambiar la temperatura de 200 a 220º C, la respuesta aumenta, *en promedio,* 1,8 unidades. Análogamente se pueden calcular los efectos principales para el tiempo (ti) y el tipo de mantequilla (M). (El efecto principal de un factor se designa con la misma notación que la utilizada para designar al propio factor).

$$\text{ti} = \frac{y_3 + y_4 + y_7 + y_8}{4} - \frac{y_1 + y_2 + y_5 + y_6}{4} = 0{,}15$$

$$\text{M} = \frac{y_5 + y_6 + y_7 + y_8}{4} - \frac{y_1 + y_2 + y_3 + y_4}{4} = 0{,}8$$

Puede observarse que los efectos principales corresponden al promedio de los valores de una cara del cubo menos el promedio de los valores en la cara opuesta, tal y como se indica en la Figura 42.3.

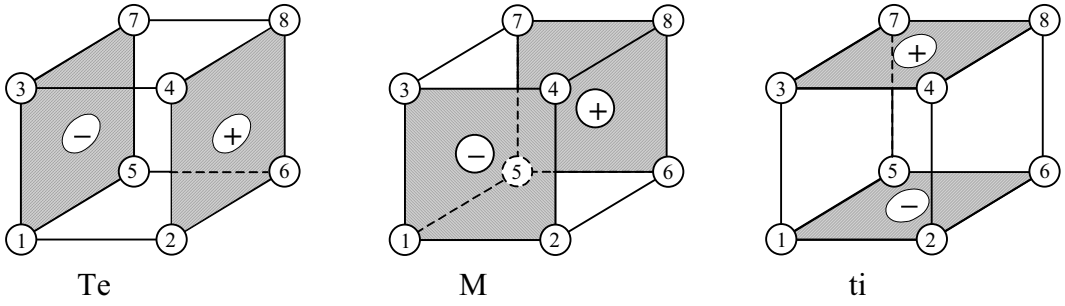


Figura 42.3. *Esquema de cálculo de los efectos principales en un diseño 2^3*

Hemos visto que el efecto principal del tiempo es 0,15, pero esto no siempre significa que cuando el tiempo pasa de 20 a 25 minutos la respuesta aumenta 0,15 unidades. Es más, en nuestro ejemplo esto no ocurre en ningún caso concreto, sino solo en promedio. Para tener una visión completa de la influencia de los factores es necesario calcular las interacciones.

Veamos si en nuestro caso interaccionan el tiempo y el tipo de mantequilla. Para ello haremos los siguientes cálculos:

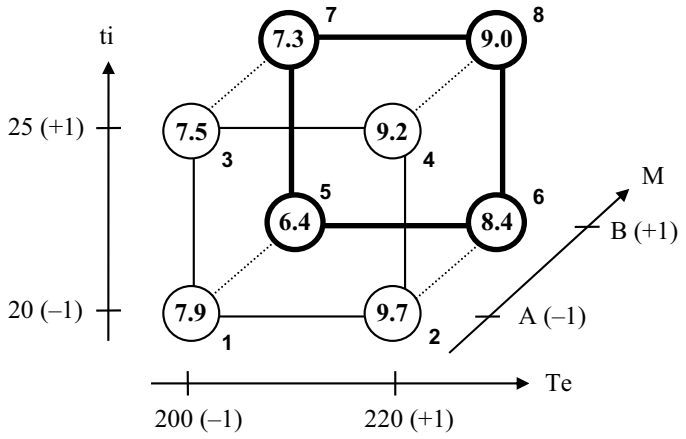
1. Efecto principal del tiempo con el tipo de mantequilla a nivel +1 (nos olvidamos de las respuestas con el tipo de mantequilla a nivel –1 o, lo que es lo mismo, nos centramos en la cara trasera del cubo).

$$\text{ti (M+)} = \frac{y_7 + y_8}{2} - \frac{y_5 + y_6}{2} = 0{,}75$$

2. Efecto principal del tiempo con la mantequilla a nivel –1 (cara delantera del cubo):

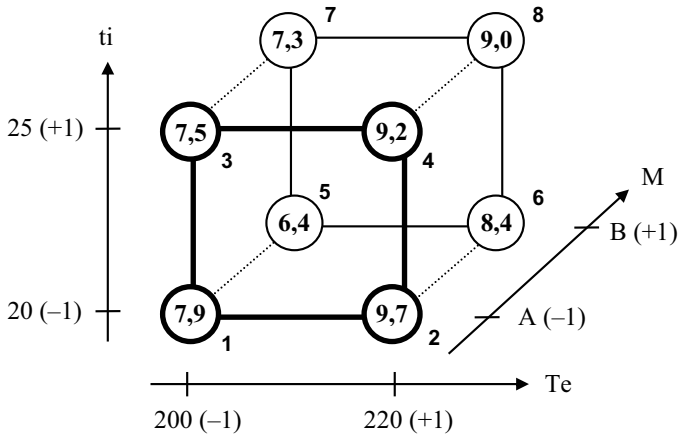
$$\text{ti (M–)} = \frac{y_3 + y_4}{2} - \frac{y_2 + y_1}{2} = -0,45$$

Por tanto, el efecto del tiempo es distinto según se use uno u otro tipo de mantequilla. Con el tipo A aumentar el tiempo de horneado hace disminuir la respuesta, mientras que con el tipo B la hace aumentar. Tiempo y mantequilla interaccionan[3] ya que el efecto de uno depende del nivel a que se encuentra el otro.



Efecto principal del tiempo considerando solo los resultados con la mantequilla **tipo B**

$$\text{ti (M+)} = \frac{y_7 + y_8}{2} - \frac{y_6 + y_5}{2} = 0,75$$



Efecto principal del tiempo considerando solo los resultados con la mantequilla **tipo A**

$$\text{ti (M-)} = \frac{y_3 + y_4}{2} - \frac{y_2 + y_1}{2} = -0,45$$

Interacción Tiempo – Tipo de mantequilla

$$\text{tiM} = \frac{1}{2}\text{ti (M+)} - \frac{1}{2}\text{ti (M-)} = \frac{y_7 + y_8 + y_1 + y_2}{4} - \frac{y_5 + y_6 + y_3 + y_4}{4} = 0,6$$

Figura 42.4. *Cálculo de la interacción de 2 factores (tiempo y tipo de mantequilla)*

[3] Aunque en rigor esta afirmación es prematura, ya que es necesario analizar si el valor obtenido es significativamente distinto de cero.

La forma de cuantificar la interacción de 2 factores podría ser simplemente la diferencia de los 2 efectos calculados, pero tiene muchas ventajas cuantificarlo como 1/2 de dicha diferencia.

$$\text{tiM} = \frac{1}{2}\left[\text{ti (M+) - ti (M-)}\right] = \frac{y_7 + y_8 + y_1 + y_2}{4} - \frac{y_5 + y_6 + y_3 + y_4}{4} = 0,6$$

De esta forma se consigue que la fórmula para el cálculo de las interacciones tenga el mismo aspecto que la del cálculo de los efectos principales: promedio de la mitad de las respuestas menos promedio de la otra mitad, aunque en este caso las mitades ya no coinciden con las caras del cubo, tal como ocurría con los efectos principales, sino con sus planos diagonales (Figura 42.5).

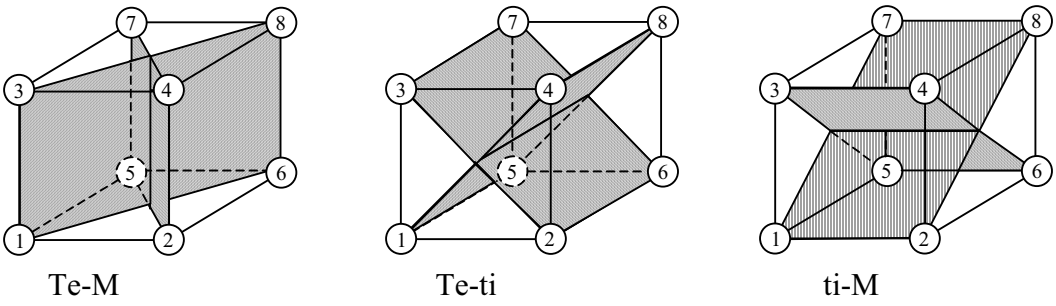


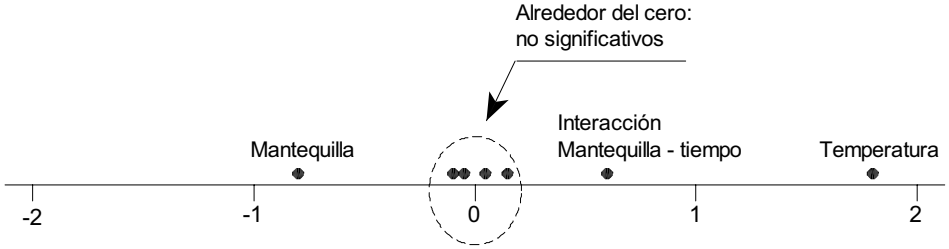
Figura 42.5. *Esquema de cálculo de las interacciones de 2 factores en un diseño 2^3.*

Como los efectos se calculan a partir de las respuestas, y estas están afectadas por el error experimental, los efectos también son variables aleatorias y el hecho de que no sean exactamente iguales a cero no significa que realmente afecten. Hace falta analizar si son significativamente distintos de cero (también llamados, simplemente, "significativos").

En nuestro caso, los efectos significativos resultan ser[4]:

Efecto	Valor
Temperatura	1,8
Tipo de Mantequilla	-0,8
Interacción Tipo de mantequilla - tiempo	0,6

[4] Los efectos son: Temperatura = 1,8; tiempo = 0,15; Mantequilla = -0,8; Interacción Temperatura – tiempo: -0,1; Interacción Temperatura – Mantequilla: 0,05; Interacción tiempo – Mantequilla = 0,6; Interacción Temperatura – tiempo – Mantequilla = -0,05. Un simple diagrama de puntos de los efectos permite identificar cuales cabe considerar como significativos.



La conclusión para la temperatura es que al pasar de 200 a 220º C la respuesta aumenta en 1,8 unidades. En el caso de la mantequilla, como interacciona con el tiempo habrá que hablar de su efecto para cada uno de los tiempos considerados. Una buena forma de hacerlo es mediante un gráfico como el de la Figura 42.6, en el que cada valor corresponde al promedio de los obtenidos para la respuesta en las condiciones que se indican. Este gráfico permite poner de manifiesto cuál es el efecto de cambiar el tipo de mantequilla para cada uno de los valores del tiempo, así como el efecto de aumentar el tiempo para cada uno de los tipos de mantequilla.

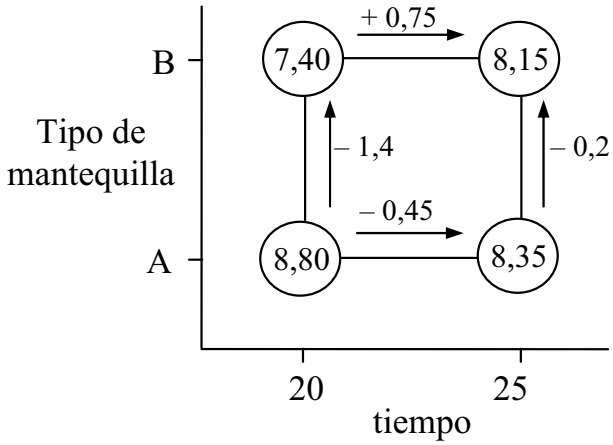


Figura 42.6. *Interacción tiempo - Tipo de mantequilla. Representación tipo A*

Hay que reconocer que hemos obtenido mucha información realizando solo 8 experimentos, y no solo hemos podido estudiar la influencia de cada uno de los factores, sino que también hemos identificado y cuantificado las interacciones entre ellos, aspecto tremendamente importante para entender cómo afectan a la respuesta e imposible de identificar moviéndolos uno a uno.

Existe otra dificultad que podría poner en jaque nuestro método, y es que el número de experimentos crece exponencialmente con el número de factores. Así, si tenemos 7 u 8 factores (cosa en principio nada rara), sería necesario realizar 128 o 256 experimentos (2^7 o 2^8 respectivamente), número que parece excesivo en muchos casos. Pero no hay que preocuparse porque también existe solución para esto. Se trata de utilizar diseños fraccionales, que con solo una parte (“fracción”) del diseño completo permiten obtener la información más relevante. Se trata, sin duda, de un tema apasionante, y sus posibilidades de aplicación, son enormes.

43 ¿Por qué funciona el algoritmo de Yates?

En primer lugar conviene aclarar qué es esto del algoritmo de Yates. Se trata de una técnica para el cálculo de los efectos en diseños factoriales a 2 niveles que aparece en los libros y que es muy usada (o quizá habría que decir "*era* muy usada", porque cada vez más los ordenadores se encargan de estas tareas). Su aplicación no requiere deducir ni razonar nada. Simplemente se hacen una serie de sumas y restas, al final una división, y, sin que se entienda muy bien cómo, ya se tienen los efectos. El algoritmo funciona igual para cualquier número de factores.

Para un diseño 2^3 se aplica de la siguiente forma: se colocan las respuestas en el orden estándar[1] de la matriz de diseño, y a continuación tantas columnas auxiliares como factores se tienen. La primera columna auxiliar se rellena a partir de las respuestas, los 4 primeros valores son las sumas de las respuestas dos a dos, y los otros 4 las diferencias (el de abajo menos el de arriba). La segunda columna auxiliar se calcula igual pero utilizando los valores de la primera columna en vez de las respuestas. La tercera columna se calcula a partir de la segunda, y así sucesivamente. Los valores de la última columna auxiliar se dividen por los situados en la columna de divisores, el primero de ellos es igual al número de condiciones experimentales y todo el resto son igual a la mitad de ese valor. La identificación de a qué efecto corresponde cada valor se hace atendiendo a dónde están situados los signos + en su fila correspondiente de la matriz de diseño. Así, si solo hay un signo + y está situado en la columna del factor A, el efecto corresponderá al efecto principal de A, si hay 2 signos + y están en las columnas A y B, tendremos la interacción AB, etc. El esquema de la Figura 43.1 ayuda a entender cómo funciona esto.

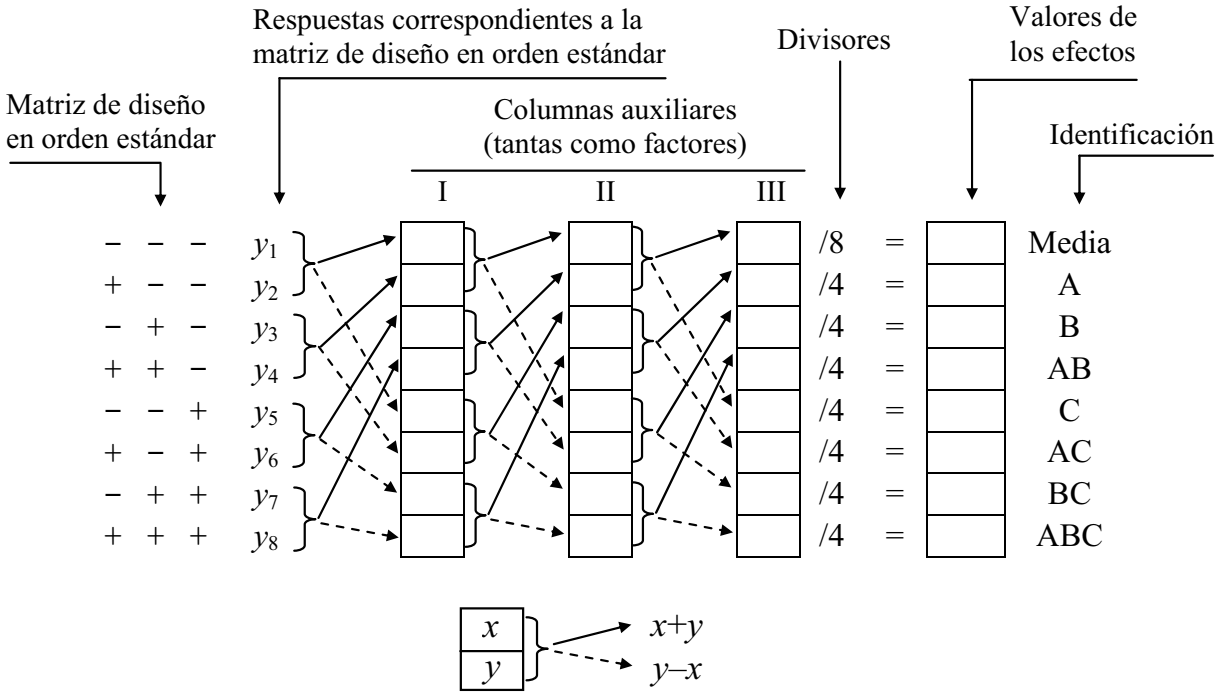


Figura 43.1. *Esquema de cálculo de los efectos en un diseño 2^3 utilizando el algoritmo de Yates*

[1] Orden estándar es el que figura en la matriz de diseño de la Figura 43.1. En la primera columna se van alternando los signos menos y más, en la segunda columna se alternan 2 menos y 2 más, en la tercera 4 y 4, si hubiera una cuarta se alternarían 8 y 8, y así sucesivamente.

Frank Yates presentó este procedimiento en un trabajo que publicó en 1937 con el título "*The Design and Analysis of Factorial Experiments*"[2]. Se trata de un excelente trabajo, de 95 páginas, donde se presentan los diseños factoriales a 2 y 3 niveles, incluyendo también comentarios para diseños con factores a más niveles, siempre ilustrados con casos prácticos de sus investigaciones agrícolas. En un apartado de la introducción, con el título "Nuevo Material", explica: "En el aspecto computacional se presenta un nuevo método para evaluar los efectos de los tratamientos con factores solo a 2 niveles (pág. 15)". Y si uno va a esa página se encuentra con el algoritmo aplicado a un diseño 2^3, sin darle mayor importancia y sin explicar ni el cómo ni el porqué se le había ocurrido hacerlo de esta forma.

Para comprobar su funcionamiento, vamos a aplicar el método que propone a unas respuestas genéricas correspondientes a diseños 2^1, 2^2 y 2^3

Tabla 43.1. *Expresiones de la última columna auxiliar al aplicar el algoritmo de Yates*

Diseño	Respuestas (orden estándar)	Última columna auxiliar	Efectos a que conducen
2^1	y_1	$y_1 + y_2$	Media
	y_2	$-y_1 + y_2$	A
2^2	y_1	$y_1 + y_2 + y_3 + y_4$	Media
	y_2	$-y_1 + y_2 - y_3 + y_4$	A
	y_3	$-y_1 - y_2 + y_3 + y_4$	B
	y_4	$y_1 - y_2 - y_3 + y_4$	AB
2^3	y_1	$y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8$	Media
	y_2	$-y_1 + y_2 - y_3 + y_4 - y_5 + y_6 - y_7 + y_8$	A
	y_3	$-y_1 - y_2 + y_3 + y_4 - y_5 - y_6 + y_7 + y_8$	B
	y_4	$y_1 - y_2 - y_3 + y_4 + y_5 - y_6 - y_7 + y_8$	AB
	y_5	$-y_1 - y_2 - y_3 - y_4 + y_5 + y_6 + y_7 + y_8$	C
	y_6	$y_1 - y_2 + y_3 - y_4 + y_5 - y_6 + y_7 - y_8$	AC
	y_7	$y_1 + y_2 - y_3 - y_4 - y_5 - y_6 + y_7 + y_8$	BC
	y_8	$-y_1 + y_2 + y_3 - y_4 + y_5 - y_6 - y_7 + y_8$	ABC

Puede comprobarse que las operaciones indicadas, a falta de la división final, conducen a los valores de los efectos que obtendríamos aplicando sus definiciones (las de efecto principal y de interacciones). Podríamos incluir en la lista los diseños 2^4 e incluso con más factores, si no fuera por problemas de espacio en el papel, y podríamos comprobar que los resultados obtenidos para los efectos siempre coinciden con los obtenidos aplicando el algoritmo de Yates.

Para entender cómo funciona hay que hacer uso del modelo subyacente para la respuesta cuando se realiza un diseño 2^k. Por ejemplo, en un diseño 2^3 lo que hacemos es explicar la respuesta a través de un modelo del tipo:

$$Y = b_0 + b_1\mathrm{A} + b_2\mathrm{B} + b_3\mathrm{C} + b_{12}\mathrm{AB} + b_{13}\mathrm{AC} + b_{23}\mathrm{BC} + b_{123}\mathrm{ABC}$$

[2] F. Yates "The Design and Analysis of Factorial Experiments". Technical Communication No. 35. Commonwealth Bureau of Soils. Harpenden, Reino Unido, 1937.

Y el cálculo de los efectos equivale al cálculo de los coeficientes del modelo[3].

Llamando **b** al vector de coeficientes, **Y** al vector de respuestas y **X** a la matriz que contiene los valores de las variables regresoras, la expresión que permite deducir el vector de coeficientes es:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

La matriz **X** contiene una primera columna con todos sus elementos iguales a 1 y a continuación la matriz del diseño junto con las columnas correspondientes a las interacciones (producto signo a signo de los factores correspondientes), es decir:

$$\mathbf{X} = \begin{array}{c} \begin{array}{cccccccc} A & B & C & AB & AC & BC & ABC \end{array} \\ \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

Al ser **X** una matriz ortogonal, **X'X** es diagonal, y es fácil comprobar que para un diseño 2^3:

$$(\mathbf{X'X})^{-1} = \begin{bmatrix} 1/8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/8 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/8 \end{bmatrix}$$

Y en general, los valores de la diagonal serán $1/2^k$. Multiplicar una matriz por otra diagonal con todos los elementos de la diagonal iguales, equivale a multiplicarla por una constante igual al valor de los elementos de la diagonal. Para los diseños factoriales, podemos poner:

$$\mathbf{b} = \frac{1}{2^k} \cdot \mathbf{X}'\mathbf{Y}$$

[3] Los valores de los efectos son el doble que los coeficientes. Esto es debido a que los coeficientes indican cuanto cambia la media de la respuesta al aumentar una unidad la variable regresora correspondiente, mientras que los efectos indican cuanto cambia la media de la respuesta al pasar la variable de –1 a +1, es decir, al saltar 2 unidades.

Vayamos ahora al algoritmo. En el caso concreto de un diseño 2^3, al pasar de la columna de respuestas a la primera columna auxiliar, y de esta a las siguientes, lo que hacemos es multiplicar por la matriz:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

De forma que en este diseño tenemos:

Columna auxiliar I: $\mathbf{A}\cdot\mathbf{Y}$
Columna auxiliar II: $\mathbf{A}\cdot(\mathbf{A}\cdot\mathbf{Y}) = \mathbf{A}^2\cdot\mathbf{Y}$
Columna auxiliar III: $\mathbf{A}\cdot[\ \mathbf{A}\cdot(\mathbf{A}\cdot\mathbf{Y})] = \mathbf{A}^3\cdot\mathbf{Y}$

Es fácil comprobar que $\mathbf{A}^3$ coincide con $\mathbf{X}'$, por lo que en la última columna auxiliar tenemos $\mathbf{X}'\mathbf{Y}$. Ya solo falta dividir por el número de condiciones de experimentación para obtener los coeficientes del modelo, o por la mitad de dicho número (excepto para el primer valor que es la media) para obtener los efectos, tal como indica el algoritmo de Yates.

44 ¿Por qué cuando se representan valores en papel probabilístico normal (ppn), en la fórmula que da la ordenada se resta 0,5 del número de orden?

La función de distribución de una Normal representada en ppn es una recta en la que a cada uno de sus puntos le corresponde un valor de la variable. Por ejemplo, si se trata de una N(0;1), al valor $X = 0$ le corresponde el punto (0; 0,5) y a $X = 1$ le corresponde (1; 0,84), ya que $P(X<1) = 0,84$.

Si en vez de representar la distribución completa representamos un conjunto de valores, en vez de la recta tendremos un conjunto de puntos. Si sabemos a que distribución pertenecen estos valores, es fácil determinar la ordenada que corresponderá a cada uno de ellos, simplemente a cada valor x le corresponderá $F(x) = P(X \leq x)$, fácil de calcular conociendo los parámetros de la distribución.

Pero lo habitual es representar los valores sin saber a qué distribución pertenecen. Lo que hacemos es representarlos como si pertenecieran a una Normal con parámetros sin determinar, y a la vista de la disposición que adoptan los puntos en el ppn juzgamos correcta o no la suposición de normalidad. Cuando representamos los efectos obtenidos en un diseño factorial, lo hacemos como si todos pertenecieran a una misma distribución. Una vez representados, observamos que algunos se alinean aproximadamente según una recta que pasa por el punto (0; 0,5), y estos son los que consideramos no significativos, ya que su disposición hace pensar que pertenecen a una Normal con media cero. Los otros, los que se alejan de la recta por los extremos, serán los que consideraremos significativos.

Si tenemos n valores, y una vez ordenados de menor a mayor, i representa su número de orden, una forma de asignar las ordenadas es haciendo $F(X_i) = i/n$. Si tenemos muchos puntos, supongamos que 10.000, esta fórmula puede ser buena, ya que una vez ordenados es razonable considerar que el que ocupa la posición 1.000 es un buen representante del que en la población deja también un 10% de valores por debajo y le

$F(x_i)= i/n$	x_i
0,1	-1,28155
0,2	-0,84162
0,3	-0,52440
0,4	-0,25335
0,5	0
0,6	0,25335
0,7	0,52440
0,8	0,84162
0,9	1,28155
1	*

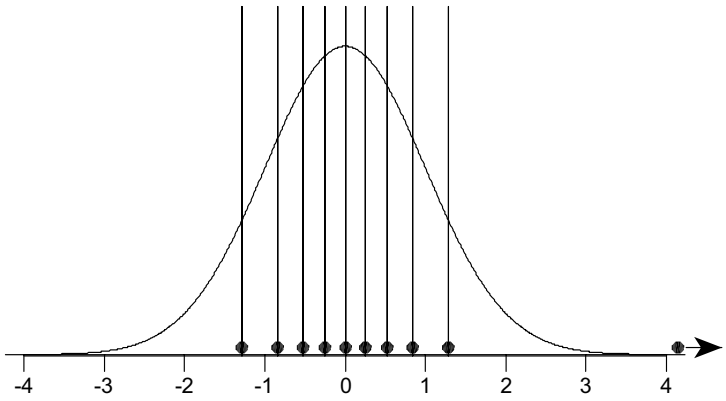


Figura 44.1. *Distribución Normal dividida en 10 zonas y puntos cuya función de distribución corresponde a las ordenadas que se asigna cuando se representan 10 valores en ppn con la expresión i/n*

podemos asignar un valor de la función de distribución de 0,10. Al que está en la posición 2.000 le asignamos 0,2, etc. Pero si vamos a representar pocos puntos, por ejemplo 10, ya no es tan razonable suponer que el menor representa al que en la población deja por debajo el 10%, el segundo al que deja por debajo al 20%,… ni el mayor de todos es el que representa al valor máximo de la distribución.

En la Figura 44.1 se ha dividido en 10 partes iguales el área de una distribución Normal estandarizada. Con la fórmula anterior a los 9 primeros valores a representar se les asignaría la ordenada de los puntos que están en los límites de separación de las zonas.

Pero es más razonable considerar que cada uno de los 10 valores representa al conjunto de cada una de las 10 zonas. Será, por tanto, más adecuado asignarles valores de la función de distribución que correspondan a puntos que representen al intervalo mejor que su valor máximo. Esto se consigue de una forma muy satisfactoria restando 0,5 al número de orden, ya que así asignamos a cada uno de nuestros valores la función de distribución que corresponde al punto que divide cada zona en 2 áreas de igual probabilidad, tal como pone de manifiesto la Figura 44.2.

$F(x_i)=(i-0,5)/n$	x_i
0,05	-1,64485
0,15	-1,03643
0,25	-0,67449
0,35	-0,38532
0,45	-0,12566
0,55	0,12566
0,65	0,38532
0,75	0,67449
0,85	1,03643
0,95	1,64485

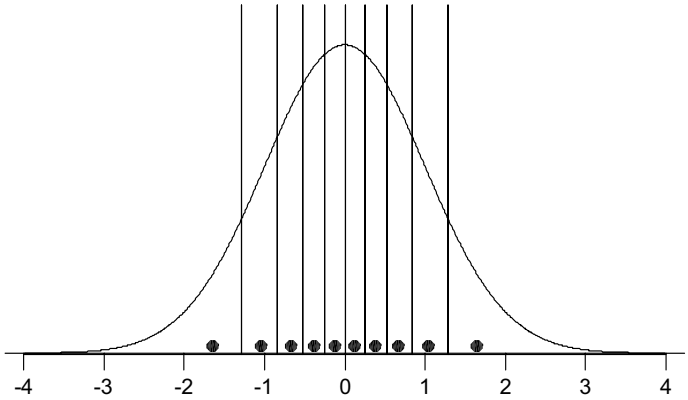


Figura 44.2. *Distribución Normal dividida en 10 zonas y puntos cuya función de distribución corresponde a las ordenadas que se asignan cuando se representan 10 valores en ppn con la expresión (i-0,5)/n*

Para ver el comportamiento de este método hemos realizado una simulación generando 10.000 muestras aleatorias de 10 elementos de una N(0; 1). Hemos ordenado cada muestra de menor a mayor y hemos calculado la media de los valores que quedan en primera posición (que son los menores de cada muestra), la media de los que quedan en segunda posición, etc. En la Tabla 44.1 están los valores medios que hemos obtenido, con los valores de la función de distribución que corresponde a estos, y también calculados con las fórmulas que antes hemos comentado.

Parece razonable asignar al valor más pequeño el valor de la función de distribución que corresponde a la media de los más pequeños, e igual para el segundo y así sucesivamente. Puede verse en la tabla que la fórmula que resta 0,5 al número de orden se acerca mucho a esta probabilidad.

Tabla 44.1. *Valores medios correspondientes a los 10 valores ordenados de una muestra aleatoria de tamaño 10 de una N(0;1), con los valores de su función de distribución y los correspondientes a las 2 fórmulas de asignación comentadas.*

Núm orden (i)	Media	P(X<Media)	i/n	(i-0,5)/n
1	-1,54	0,06	0,1	0,05
2	-1,00	0,16	0,2	0,15
3	-0,65	0,26	0,3	0,25
4	-0,37	0,36	0,4	0,35
5	-0,12	0,45	0,5	0,45
6	0,13	0,55	0,6	0,55
7	0,38	0,65	0,7	0,65
8	0,66	0,75	0,8	0,75
9	1,00	0,84	0,9	0,85
10 = n	1,54	0,94	1	0,95

Pero hay otras formas de asignar las ordenadas. Minitab, por defecto, utiliza $(i-3/8)/(n+1/4)$, que se acerca todavía más a los valores de la función de distribución de las medias de cada intervalo. Nuestra fórmula, que consiste en restar 0,5 al número de orden y dividir por el número de datos, es más sencilla y suficientemente aproximada a efectos prácticos, y por ello la más utilizada. Es la que se recomienda en textos muy conocidos como los de D.C. Montgomery o el de Box, Hunter y Hunter para representar los efectos de un diseño factorial en papel probabilístico normal.

45 En los diseños factoriales, ¿cómo se puede escribir una ecuación para la respuesta a partir de los efectos?

Lo veremos a través de un ejemplo. Supongamos que tenemos un circuito como el de la Figura 45.1 y deseamos conocer cuál es la intensidad (I) que circula por el mismo en función del valor de la resistencia (R) y de la fuente de alimentación (V).

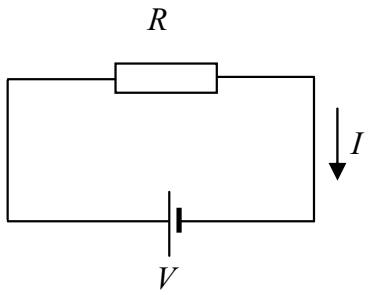


Figura 45.1: *Circuito con fuente de alimentación y resistencia*

En un caso como este la relación es perfectamente conocida, se trata de la ley de Ohm, que puede escribirse de la forma:

$$I = \frac{V}{R}$$

Supongamos que no conocemos esta relación y queremos deducirla a través de la experimentación[1]. Consideremos que nuestra zona de interés se encuentra en valores de entre 6 y 9 voltios para el voltaje y de entre 2 y 4 ohmios para la resistencia. Supongamos también que entre estos valores se considera razonable la aproximación lineal. En este caso bastará con realizar un diseño 2^2 en el que, ignorando la existencia del error experimental, se obtendrían los resultados que se indican a continuación:

V	R	I
6	2	3,00
9	2	4,50
6	4	1,50
9	4	2,25

Calculando los efectos se obtienen los siguientes resultados[2]:

	Media	V	R	VR	I
	+	–	–	+	3,00
	+	+	–	–	4,50
	+	–	+	–	1,50
	+	+	+	+	2,25
Divisor	4	2	2	2	
Efecto	2,8125	1,125	–1,875	–0,375	

[1] En general, mediante la experimentación no se pretende deducir la relación funcional exacta, sino una aproximación que resulte útil para cumplir los objetivos que se hayan planteado.

[2] Se aplica el algoritmo de los signos, que consiste en escribir la matriz del diseño ampliada con las columnas de las interacciones. Los efectos son la suma de las respuestas con los signos de su columna correspondiente, divido por la mitad del número de condiciones experimentales.

Los efectos indican cuánto cambia en promedio la respuesta al pasar el factor de nivel –1 a nivel +1, por tanto, indican el cambio al aumentar en 2 unidades el valor del factor. Sin embargo, los coeficientes de una ecuación de regresión indican cuanto cambia la respuesta en promedio al aumentar en una unidad la variable regresora.

La ecuación para explicar el comportamiento de la respuesta se escribe colocando como constante su valor medio, y corrigiendo esa media general con el valor de los factores (el subíndice C indica que están codificados, usando –1 y +1, no sus valores reales) acompañados de unos coeficientes que son los valores de los efectos divididos por 2. En nuestro caso la ecuación tendrá la forma:

$$I = \text{media} + \frac{\text{Efecto ppal. de } V}{2} V_C + \frac{\text{Efecto ppal. de } R}{2} R_C + \frac{\text{Interacción } VR}{2} V_C \times R_C$$

Y sustituyendo queda:

$$I = 2.8125 + 0.5625\, V_C - 0.9375\, R_C - 0.1875\, V_C R_C$$

Pero, tal y como se ha comentado, en esta ecuación los valores de V y R deben introducirse codificados[3], ya que al calcular los efectos se ha considerado que los valores tanto para V como R eran –1 y +1, sin tomar en consideración a qué valores (en voltios y en ohmios) corresponden esos niveles.

Así, si queremos saber cuál será el valor de la intensidad para V = 9 V y R = 2 Ω, deberemos sustituir V por +1 (9 V es el nivel + de esta variable) y R por –1 (2Ω es el nivel – para R), obteniendo:

$$I = 2{,}8125 + 0{,}5625\ (+1) - 0{,}9375\ (-1) - 0{,}1875\ (+1)(-1)$$

$$I = 4{,}5$$

Para tener la ecuación en las unidades originales de V y R es necesario decodificar estas variables, para ello puede usarse la expresión:

$$X_C = -1 + 2\ \left(\frac{X - X_{(-)}}{X_{(+)} - X_{(-)}} \right)$$

Siendo X el valor de la variable sin codificar y $X_{(+)}$, $X_{(-)}$ sus niveles alto y bajo respectivamente. Aplicando esta transformación a nuestro modelo, tenemos:

$$I = 2.8125 + 0.5625\left[-1 + 2\ \left(\frac{V-6}{3}\right)\right] - 0.9375\left[-1 + 2\ \left(\frac{R-2}{2}\right)\right] - 0.1875\left[-1 + 2\ \left(\frac{V-6}{3}\right)\right]\left[-1 + 2\ \left(\frac{R-2}{2}\right)\right]$$

[3] Es decir: +1 para el valor máximo del rango de variación, –1 para el valor mínimo, 0 para el valor medio, etc.

y operando se llega a:

$$I = \frac{3}{4}V - \frac{1}{8}VR$$

El modelo es una aproximación a la realidad. La Figura 45.2 compara la superficie de respuesta que corresponde a la relación real (curvada) junto con la obtenida a través del diseño 2^2 (plana).

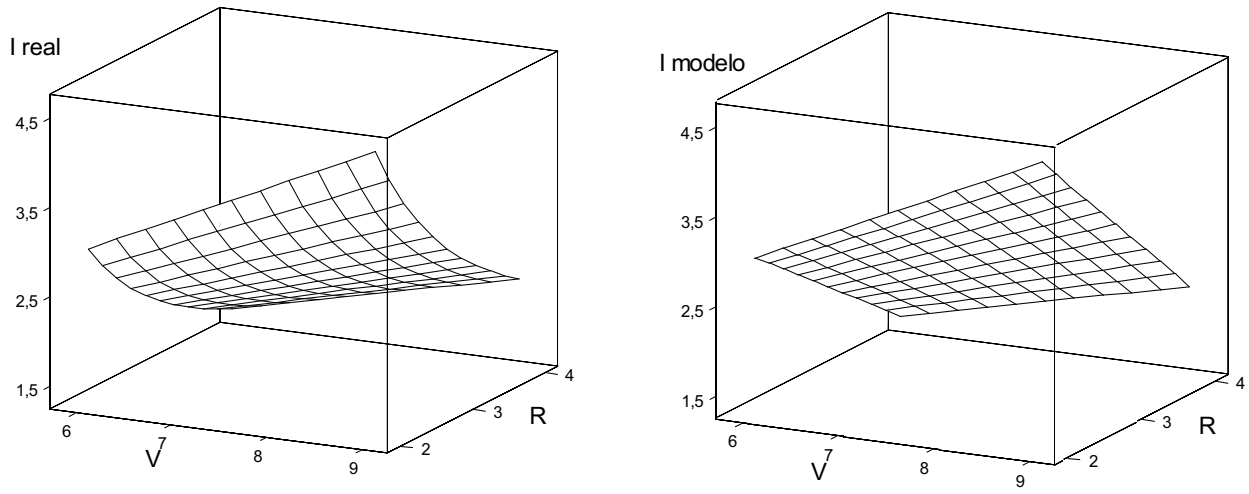


Figura 45.2. *Superficie de respuesta real y aproximación deducida a través de la experimentación. En las esquinas (puntos de experimentación) los valores coinciden*

¿Qué ocurriría si en vez de la resistencia se mide la conductancia (su inversa: $G = 1/R$)? Ahora la fórmula es $I = V \cdot G$, y la tabla de resultados de la experimentación será:

V	*G*	*I*
6	0,25	1,50
9	0,25	2,25
6	0,50	3,00
9	0,50	4,50

Obteniéndose los efectos:

	Media	*V*	*R*	*VR*	*I*
	+	–	–	+	1,50
	+	+	–	–	2,25
	+	–	+	–	3,00
	+	+	+	+	4,50
Divisor	4	2	2	2	
Efecto	2,8125	1,125	1,875	0,375	

Y a partir de los efectos podemos escribir el modelo:

$$I = 2{,}8125 + 0{,}5625\ V_C + 0{,}9375\ R_C + 0{,}1875\ V_C R_C$$

Siendo ahora la expresión con las variables originales:

$$I = 2,8125 + 0,5625\left[-1+2\left(\frac{V-6}{3}\right)\right] + 0,9375\left[-1+2\left(\frac{G-0,25}{0,25}\right)\right] + 0,1875\left[-1+2\left(\frac{V-6}{3}\right)\right]\left[-1+2\left(\frac{G-0,25}{0,25}\right)\right]$$

Y esta expresión, después de operar y simplificar se reduce a[4]:

$$I = VG$$

En este caso, como la relación verdadera es un polinomio sin términos cuadráticos, el modelo obtenido coincide perfectamente con la realidad. Es interesante observar como la selección de una métrica adecuada para los factores (medir la conductancia en vez de la resistencia) influye de forma importante en la bondad del modelo obtenido.

Un comentario final: presentar los resultados en forma de modelo matemático y utilizarlo para deducir los valores óptimos de los factores puede resultar muy útil, pero conviene no dejarse llevar por las apariencias dándole una generalidad y un poder de predicción que no tiene. La información que contiene el modelo es exactamente la misma que la que aportan los efectos junto con la media de los resultados.

---

[4] Por si desea repasar las operaciones:

$$\left[-1+2\left(\frac{V-6}{3}\right)\right] = -1+\frac{2V}{3}-4 = -5+\frac{2V}{3}$$

$$\left[-1+2\left(\frac{G-0,25}{0,25}\right)\right] = -1+8G-2 = -3+8G$$

$$\left[-1+2\left(\frac{V-6}{3}\right)\right]\cdot\left[-1+2\left(\frac{G-0,25}{0,25}\right)\right] = \left(-5+\frac{2V}{3}\right)\cdot\left(-3+8G\right)$$

$$\left(-5+\frac{2V}{3}\right)\cdot\left(-3+8G\right) = 15-40G-2V+\frac{16}{3}VG$$

$$0,5625\left(-5+\frac{2V}{3}\right) = -2,8125+0,375V$$

$$0,9375\left(-3+8G\right) = -2,8125+7,5G$$

$$0,1875\left(15-40G-2V+\frac{16}{3}VG\right) = 2,8125-7,5G-0,375V+VG$$

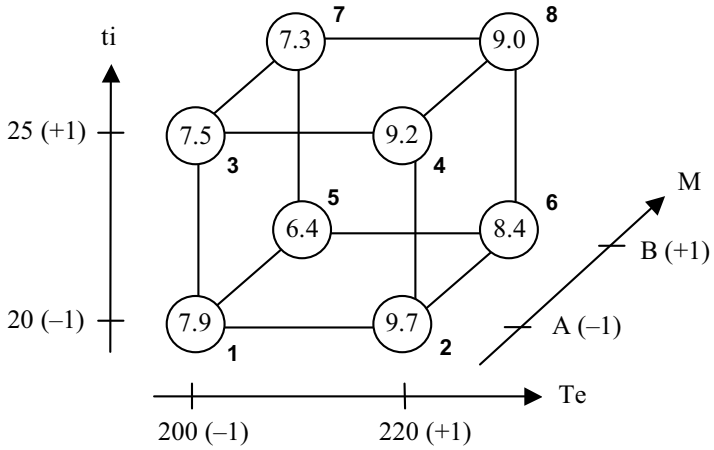
$$I = 2,8125-2,8125+0,375V-2,8125+7,5G+2,8125-7,5G-0,375V+VG$$

$$I = VG$$

46 ¿Qué es un diseño bloqueado? ¿Por qué en estos diseños no se tienen en cuenta las interacciones entre los factores de bloqueo y el resto de factores? ¿Qué ocurre si esas interacciones existen?

Cuando se lleva a cabo un plan de experimentación, lo deseable es que de un experimento a otro solo varíen aquellos factores cuyo efecto se desea estudiar. Pero lamentablemente en algunas ocasiones no se pueden realizar todos los experimentos en condiciones homogéneas, ya sea porque hay que realizarlos en diferentes periodos de tiempo (4 experimentos este sábado y otros 4 dentro de 3 semanas, por ejemplo), o porque es necesario utilizar materias primas de 2 orígenes distintos, o varias máquinas, u operarios, etc. En estos casos, la variación de las condiciones ambientales, la materia prima, los operarios o las máquinas quizá influyan en la respuesta y esta influencia podría confundirnos a la hora de sacar conclusiones.

La solución que se da en estos casos es realizar los experimentos de forma que la influencia de ese factor perturbador (a veces son varios) no afecte a la estimación de los efectos que interesan. Por ejemplo, se realiza un diseño 2^3 para determinar cuáles son las condiciones de elaboración que maximizan lo crujientes que resultan unas galletas. Los factores son la temperatura del horno (Te), el tiempo de horneado (ti) y el tipo de mantequilla (M). Supongamos que los 8 experimentos se hacen en condiciones homogéneas, obteniéndose los resultados y los efectos que se indican en la Figura 46.1.



Efectos:

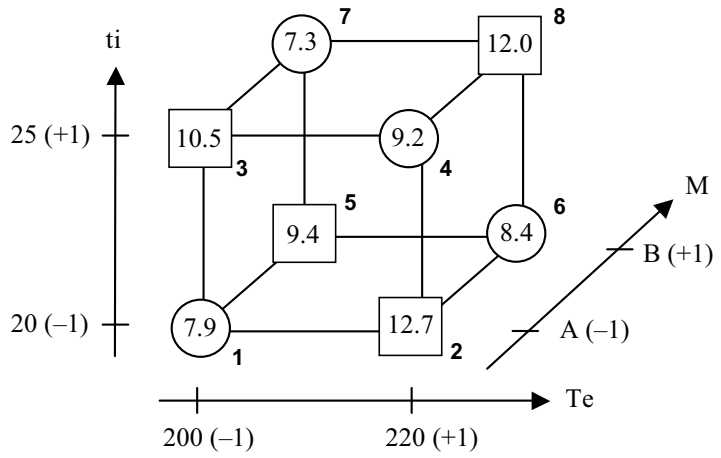
Te	1,80
ti	0,15
M	-0,80
Te*ti	-0,10
Te*M	0,05
ti*M	0,60
Te*ti*M	-0,05

Figura 46.1. *Plan de experimentación llevado cabo en condiciones homogéneas, junto con los efectos obtenidos*

Supongamos ahora que no pueden realizarse los 8 experimentos el mismo día, sino solo 4 en un día y otros 4 pasado un cierto periodo de tiempo y sospechamos que las condiciones ambientales (podría ser la humedad del ambiente, por ejemplo) pueden tener influencia en la respuesta medida. Una opción posible para repartir los 8 experimentos en los 2 días podría ser realizar primero los 4 correspondientes a la mantequilla a nivel + (los de la cara de atrás del cubo) y después los 4 con la mantequilla a nivel – (los de la cara de delante). Esta sería claramente una mala opción, ya que si el

cambio de día afecta a la respuesta, esta influencia estará confundida con el efecto de la mantequilla y podríamos sacar una conclusión errónea respecto a la influencia de este factor sobre la respuesta.

Una opción mejor sería realizar el reparto tal como se indica en la Figura 46.2, en la que los valores en un círculo corresponden a los resultados obtenidos el día 1 (primer bloque), y los que están en un recuadro corresponden al día 2 (segundo bloque). Los resultados son iguales que en el caso anterior excepto que los correspondientes al día 2 dan una respuesta 3 unidades superior a la de antes. ¿Cómo cambiará en este caso nuestra interpretación de la influencia de los factores sobre la respuesta? Pues para lo que realmente interesa estudiar, que son los efectos principales y las interacciones de 2 factores, las conclusiones son idénticas. Hemos repartido los experimentos de forma que la influencia del cambio de día queda neutraliza a la hora de estimar los efectos que nos interesan[1].



Efectos:

Te	1,80
ti	0,15
M	-0,80
Te*ti	-0,10
Te*M	0,05
ti*M	0,60
Te*ti*M	-2,95

Figura 46.2. *Plan de experimentación realizado durante 2 días con distintas condiciones ambientales. Las respuestas situadas en un círculo corresponden al día 1, y las que están en un cuadrado al día 2*

Decimos que este es un diseño bloqueado, un diseño en el que para evitar la posible influencia de factores ajenos a los sometidos a estudio, su ejecución se ha dividido en partes (bloques) dentro de las cuales las condiciones sí son homogéneas, y además esas partes han sido elegidas de forma que el efecto de los factores ajenos a la experimentación no "contaminará" la estimación de aquellos que consideremos más importantes.

Aclarado ya lo que significa esto de los diseños bloqueados y conscientes de la importancia de detectar las interacciones entre los factores, surge la pregunta, ¿Por qué en el análisis de los diseños bloqueados no se tienen en cuentas las posibles interacciones de los factores de bloqueo con los factores en estudio? La razón es básicamente terminológica: cuando un factor lo tratamos como factor de bloque,

[1] La influencia del "factor día" está confundida con la interacción de los 3 factores. Normalmente esta interacción no es significativa, por lo que haciéndolo de esta forma, lo que esperamos encontrar en esta interacción es en realidad el efecto del cambio de día.

suponemos que no interacciona con los factores en estudio, tiene lo que llamamos efecto aditivo, es decir, sube o baja globalmente el nivel de la respuesta (por el hecho de hacerse el día 2 la respuesta sube 3 unidades). Esta es una hipótesis bastante razonable cuando los factores de bloqueo son del tipo de los descritos anteriormente: tiempo, materias primas, máquinas,...

Si se considera que un factor que en principio se pensaba plantear como de bloqueo, puede interaccionar con los factores en estudio, ya no le llamaremos factor de bloqueo. En este caso se tratará, simplemente, de un factor más.

El hecho de que el marco de análisis establecido esté orientado a que los factores de bloqueo no interaccionen, puede hacer que en algunas ocasiones ni nos planteemos la posibilidad de tal interacción, como si al clasificar un factor como de bloqueo ya le quitáramos la capacidad de interaccionar. Y esto, evidentemente, no es así. La hipótesis sobre el efecto aditivo de nuestro factor de bloqueo debe venir razonada a partir de un análisis técnico del sistema con el que se experimenta, no en planteamientos generales que pueden ser ciertos en muchos casos, pero no necesariamente en el nuestro.

¿Y si es más razonable plantearse la existencia de la interacción, y después se comprueba que realmente existe? ¿Hemos tenido mala suerte? En absoluto, más bien todo lo contrario. Muchas veces, los factores por los que se bloquea son factores que afectan a las características del producto y sin embargo no se pueden mantener constantes (varían las condiciones ambientales, las máquinas, la materia prima,…) Son lo que llamamos “factores de ruido”. Si el factor de bloqueo es también un factor de ruido e interacciona con un factor de control, significa que el efecto de uno depende del valor del otro, y podremos dar al factor de control el valor que minimice la influencia del factor de ruido.

Por ejemplo, si en nuestro caso de las galletas descubrimos que la humedad interacciona con el tipo de mantequilla, esto significa que el efecto de la humedad en lo crujientes que resultan las galletas depende del tipo de mantequilla que contengan, y podremos elegir el que minimice la influencia de la humedad. En definitiva, la interacción nos da la oportunidad de conseguir un objetivo siempre deseado, como es disminuir la variabilidad en las características del producto.

47 ¿Por qué es razonable suponer no significativas las interacciones de 3 o más factores?

En primer lugar vamos a intentar ver geométricamente lo que significa una interacción de 2 factores. La Figura 47.1 muestra la superficie $Z = 1 + X + Y$. Si consideramos que X e Y son los factores y Z la respuesta, esta será una situación en que los factores no interaccionan, es decir, el efecto de X sobre la respuesta no depende del valor de Y, (ni el efecto de Y depende de X). En concreto, aumentar una unidad el valor de X aumenta también una unidad el valor de Z, independientemente del valor de Y. La parte derecha de la Figura 47.1 (es el cubo de la izquierda visto por la cara X) muestra claramente cómo el efecto de X no depende de Y (las pendientes de todas las rectas son iguales).

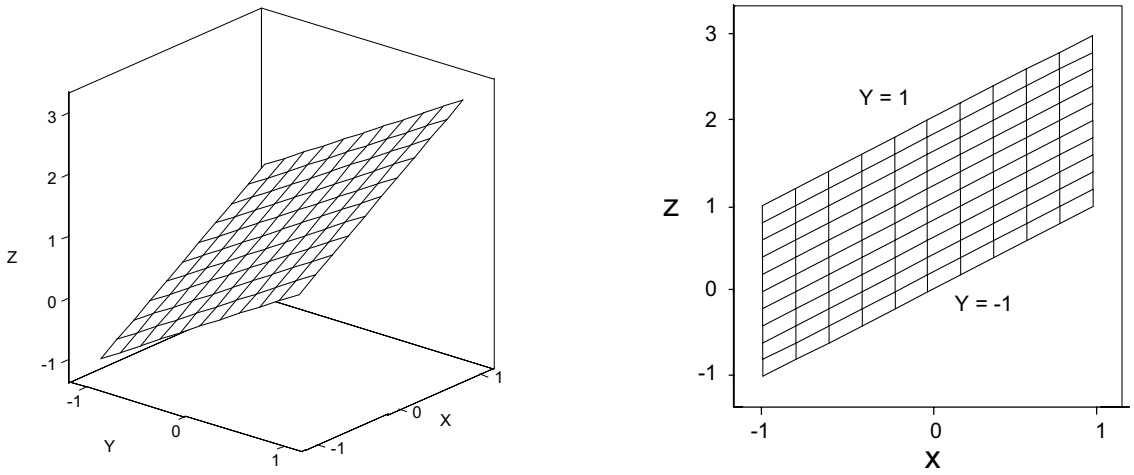


Figura 47.1. *Superficie de respuesta para la ecuación Z = 1 + X + Y*

Veamos ahora la Figura 47.2, la ecuación de la superficie representada es $Z = 1{,}5 + 2X + Y + 1{,}5XY$. En este caso X e Y sí que interaccionan, ya que el efecto de X sobre la respuesta depende del valor de Y. Al pasar X de -1 a +1 crece más la respuesta cuando Y toma el valor +1 que cuando toma el –1.

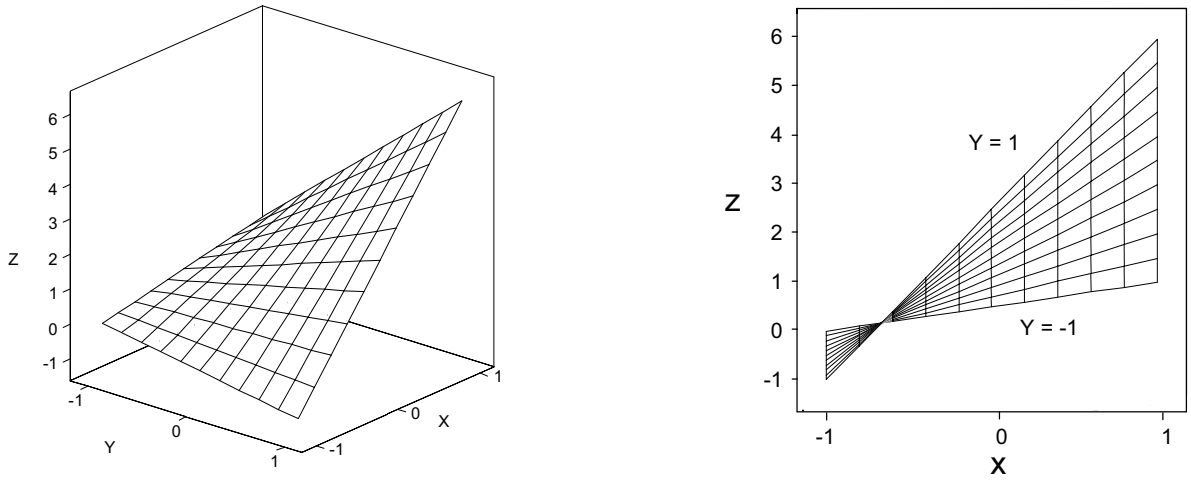


Figura 47.2. *Superficie de respuesta para la ecuación Z = 1,5 + 2X + Y + 1,5XY*

Obsérvese que la existencia o no de interacción está relacionada con lo "retorcida" que es la superficie. Lo que ocurre cuando se tiene una interacción de 3 factores no es fácil representarlo gráficamente (3 factores más la respuesta exigirían 4 dimensiones), pero la idea general es que cuanto mayor es el grado de las interacciones que hacen falta para describir una superficie, más irregular o "sofisticada" resulta ser.

Cuando en la práctica se explora una superficie de respuesta, no se espera que sea muy irregular, por las siguientes razones:

1. Las superficies que responden a fenómenos físicos no son superficies caprichosas, sino que en general responden a unos ciertos tipos conocidos[1] que presentan suavidad y continuidad (se excluyen los casos en que hayan cambios de estado en la región de experimentación).

2. Siempre se experimenta en un rango limitado de variación de los factores. Así, aunque la superficie pueda tener cierta complicación, es muy probable que en el reducido rango en el que se está experimentando, pueda ser explicada por una ecuación sencilla.

3. Siempre se conoce algo de la superficie que se explora y, en muchos casos se puede tener una certeza de que no va a ser necesario incluir interacciones de 3 o más factores.

De hecho, si no fuera por estas razones, la experimentación sería una técnica poco prometedora, ya que, como se dice en el texto de Box, Hunter y Hunter: "... Esto es una suerte, ya que de lo contrario, hacer un mapa de una superficie llena de estalagmitas, o parecida al lomo de un erizo sería totalmente imposible con un número razonable de experimentos".

Por otra parte, mediante los diseños factoriales a 2 niveles, lo que hacemos es buscar una aproximación a la relación funcional que liga los factores con la respuesta, del tipo de la que se obtendría al descomponer una función en serie de Taylor. No es que el modelo obtenido a través de la experimentación sea la descomposición en serie de Taylor de la función buscada, sino que es una aproximación del mismo tipo. Por ejemplo, en la pregunta sobre cómo se puede escribir un modelo para la respuesta a partir de los efectos, experimentábamos con un circuito en el que el voltaje (V) tomaba valores de 6 y 9 V, y la resistencia (R) de 2 y 4 Ω, llegando al modelo para la intensidad (I):

$$I = \frac{3}{4}V - \frac{1}{8}VR$$

Descomponiendo la función real ($I = V/R$) en serie de Taylor en torno al punto central de la región de experimentación ($V = 7{,}5$ y $R = 3$), incluyendo los términos de primer orden y el de derivadas cruzadas se obtiene la función:

$$I = \frac{2}{3}V - \frac{1}{9}VR$$

[1] Tipo montaña, teja plana, teja inclinada, silla de montar, ...

Ambas ecuaciones son del mismo tipo (efecto principal de *V* e interacción *VR*), en el modelo deducido a partir de la experimentación, la superficie se apoya en las 4 esquinas de la superficie real (que es la curvada), mientras que en el caso de la descomposición en serie de Taylor la superficie es tangente a la superficie real en su punto central (Figura 47.3).

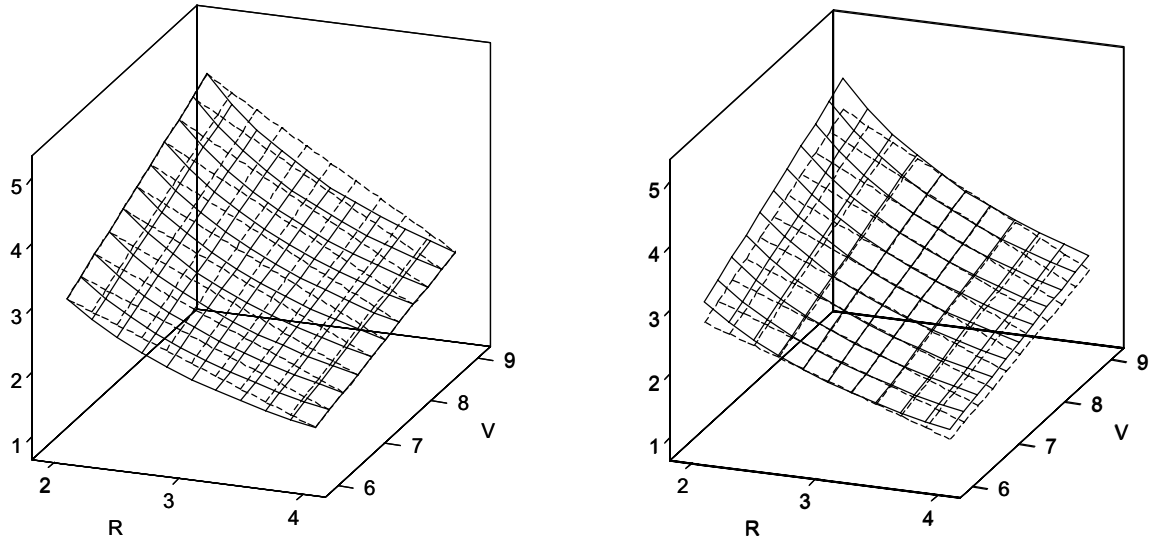


Figura 47.3. *Aproximación a la función I=V/R deducida a través de la experimentación (izquierda) y por descomposición en serie de Taylor (derecha). La función real está en trazo continuo y las aproximaciones a trazos*

Si estamos ante una función de las que esperamos encontrarnos en la práctica (suave, regular, sin grandes altibajos), la zona de experimentación, que se limitará a un rango reducido de variación de los factores, se podrá aproximar razonablemente bien descomponiéndola en serie de Taylor hasta el término de las derivadas cruzadas de segundo orden, sin necesidad de utilizar los términos correspondientes a las derivadas cruzadas de tercer orden (equivalente a las interacciones de 3 factores) y menos aún términos de cuarto orden (interacciones de 4) o superiores.

Tampoco es que se dogmatice la no existencia de interacciones de 3 o más factores, pero como el número de experimentos a realizar siempre resulta escaso, vale la pena ser consciente de la importancia de cada uno de los efectos que es posible estimar y dedicar los esfuerzos solo a aquellos que mejor ayudarán a explicar el comportamiento de la respuesta. Por suerte, las interacciones de 3 o más factores no suelen ser importantes.

48 ¿Qué hacer si al aleatorizar el orden de experimentación se obtiene el orden estándar de la matriz de diseño?

Siempre se insiste en que una cosa es el orden en que se presenta la matriz de diseño (relación de las condiciones en que se va a experimentar) y otra el orden en que se deben realizar los experimentos. La matriz de diseño se presenta en el llamado "orden estándar" (Figura 48.1) tanto por sus ventajas a la hora de escribirlo (la regla de construcción es muy fácil) como por la comodidad de tener un orden establecido al que nos podemos referir sin tener que detallar cuáles son las condiciones una a una.

Orden	Matriz de diseño			Respuestas
	A	*B*	*C*	
1	–	–	–	y_1
2	+	–	–	y_2
3	–	+	–	y_3
4	+	+	–	y_4
5	–	–	+	y_5
6	+	–	+	y_6
7	–	+	+	y_7
8	+	+	+	y_8

Figura 48.1. *Matriz de un diseño 2^3 en orden estándar*

Pero realizar los experimentos en este orden tiene el inconveniente de que si la respuesta tiende a aumentar (o a disminuir) a medida que se van realizando[1], como los efectos principales se calculan haciendo el promedio de respuestas con el factor a nivel + menos el promedio con el factor a nivel –, el efecto principal del factor situado en la última columna (el C en el caso de un 2^3) será el promedio de la mitad de experimentos realizados al final menos la mitad de los realizados al principio, y al hacerlo de esta forma la influencia del orden puede llevarnos a conclusiones equivocadas. Para evitar este tipo de situaciones, aleatorizamos el orden de realización de los experimentos, y si el orden tuviera alguna influencia, esperamos que esta se difumine entre todos los efectos sin concentrarse en ninguno de forma especial.

Planteada la razón de la aleatorización surge la pregunta de qué ocurre si al aleatorizar el orden obtenido resulta ser el estándar. La respuesta es polémica. Vamos a comentarla bajo dos puntos de vista.

Algunas personas (les llamaremos "académicos") consideran que el orden estándar es un orden como cualquier otro y no debería haber ninguna razón para rechazarlo. Si se tienen razones para sospechar que el orden afecta a la respuesta, lo que hay que hacer es bloquear el experimento y/o tomar las medias adecuadas para que esa influencia no se presente. Además, esa influencia puede seguir cualquier patrón, y por tanto tampoco está justificado temerle más a unos (tendencias crecientes o decrecientes) que a otros.

[1] Ya sea por razones vinculadas al sistema de medida (el aparato se va descentrando...), a la materia prima (se va secando...), o a la maquinaria (se va ensuciando ...), o a lo que sea.

Otras personas (les llamaremos "prácticos") opinan que si al aleatorizar sale el orden estándar lo mejor es volver a realizar la aleatorización. Porque, aunque es verdad que la tendencia vinculada al orden puede ser de cualquier tipo, en el caso de que exista es más razonable considerar que sea monótona creciente o decreciente y en ese caso el orden estándar sale especialmente perjudicado. Además, es difícil explicar que el denostado orden estándar para realizar la experimentación es el que finalmente hay que llevar cabo[2].

Quizá en esta polémica conviene tener presente que la palabra "aleatorizar" se utiliza en estadística en dos contextos y con dos significados distintos. Una cosa es aleatorizar la forma de seleccionar los elementos de una muestra que se va a considerar aleatoria, en cuyo caso la aleatorización es una condición crítica y hay que andarse con cuidado con los relajamientos ante esta exigencia, y otra cosa es aleatorizar el orden de realización de los experimentos. Aunque la palabra es la misma (quizá la más sacralizada de las que usamos en estadística) en el último caso se trata de una práctica del tipo "curarse en salud" de forma que su significado e implicaciones son distintas.

[2] Consuela saber que las probabilidades de que esto ocurra son pocas. Exactamente 1 entre 40.320 en el caso de que aleatorice el orden de 8 experimentos, y del orden de 1 entre 21 billones si aleatoriza 16.

Estudios de capacidad y control estadístico de procesos

49 ¿Qué diferencia hay entre un estudio de capacidad a corto y largo plazo? ¿Cómo se estima la variabilidad en uno y otro caso?

El estudio de capacidad a corto plazo se refiere, tal y como su nombre indica, a la variabilidad observada cuando se considera un corto periodo de tiempo, sin dar lugar a que intervengan las causas de variabilidad que inevitablemente van apareciendo, tales como cambios en las materias primas, ensuciamiento de la máquina, cambio de operario, etc. Esta variabilidad a corto plazo se puede entender como la mínima a que se podría aspirar si se llegaran a eliminar las causas relacionadas con cambios en las condiciones de producción.

Capacidad a largo plazo es la que mide la variabilidad que realmente se produce a efectos prácticos, la que incorpora todas las causas que van apareciendo a lo largo del tiempo.

Una forma de valorar la capacidad a corto y largo plazo es ir tomando muestras cada cierto tiempo, durante un periodo lo suficientemente grande como para que hayan aparecido todas las causas que introducen variabilidad. Por ejemplo, tomar una muestra de 4 observaciones cada 30 minutos durante 3 jornadas. La variabilidad dentro de las muestras es una medida de la variación a corto plazo, y la variación global, de todos los valores, será la medida de la variación a largo plazo.

Vamos a verlo con un caso muy simplificado y con números sencillos. Supongamos que se toman 3 muestras de 4 observaciones cada una y que los valores son[1]:

Muestra	Valores			
1	2,	4,	5,	6
2	12,	13,	14,	15
3	6,	7,	8,	10

Estimación de la varianza para el estudio de capacidad a corto plazo

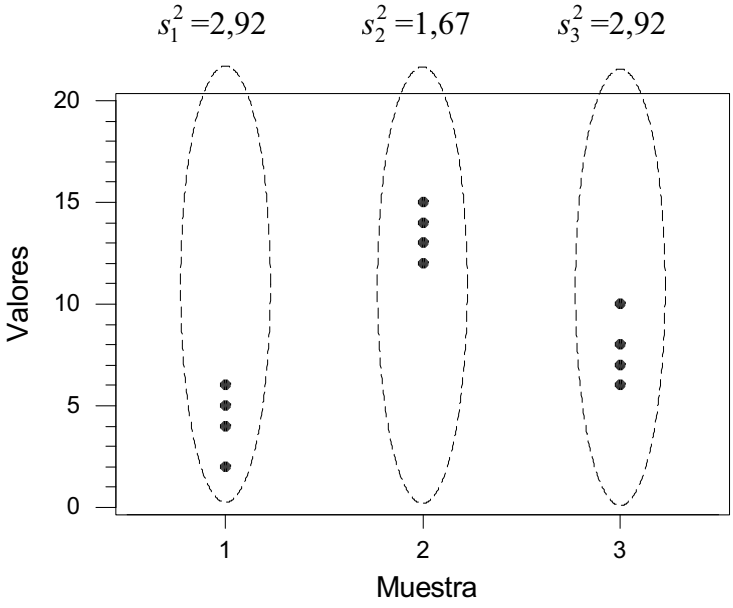


Figura 49.1. *Estimación de la varianza "dentro" ("within") de cada muestra*

[1] Para un estudio de capacidad real se tomarían más muestras. Este es un ejemplo simplificado sólo para ilustrar como se hacen los cálculos.

El mejor estimador de la varianza dentro de las muestras es la media de las varianzas en cada una de ellas (porque los tamaños de las muestras son iguales). En nuestro ejemplo: $s_R^2 = (2,92 + 1,67 + 2,92)/3 = 2,50$, y por tanto, $s_R = \sqrt{2,50} = 1,58$.

Estimación de la varianza para el estudio de capacidad a largo plazo

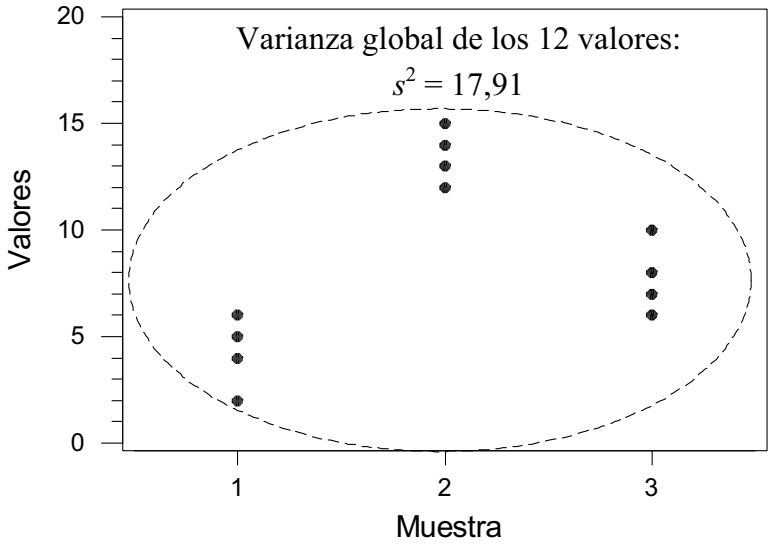


Figura 49.2. *Estimación de la varianza global ("overall")*

La varianza de todos los datos conjuntamente es $s^2 = 17,91$ y la desviación tipo: $s = \sqrt{17,91} = 4,23$

Si utilizamos un paquete de software estadístico como Minitab para realizar este estudio, colocando como tolerancias 0 y 15 (son valores arbitrarios para este ejemplo), se obtiene:

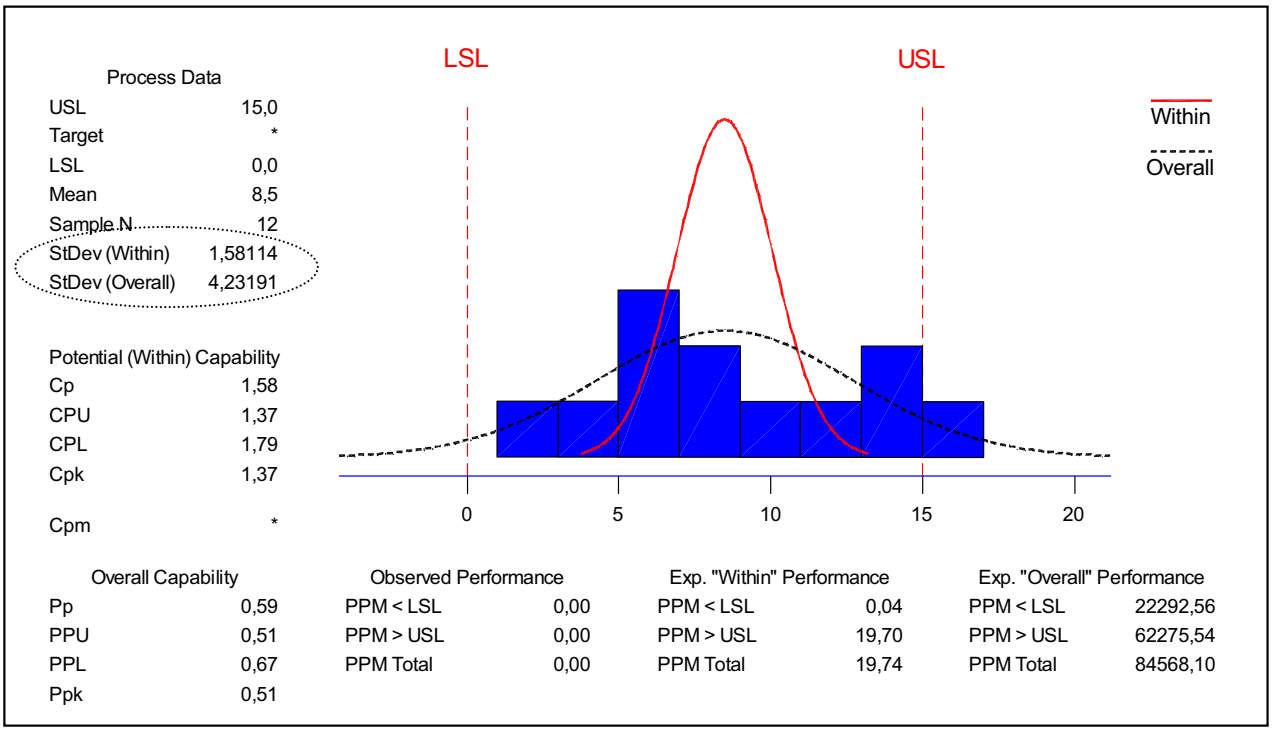


Figura 49.3. *Salida de Minitab al realizar un estudio de capacidad*

Within corresponde a la variabilidad dentro de las muestras, y *Overall* a la variabilidad global.

Antes de terminar debemos aclarar que las desviaciones tipo coinciden con las que habíamos calculado porque hemos modificado las opciones que Minitab tiene por defecto. La razón es que aunque $E(s^2) = \sigma^2$, resulta que $E(s) \neq \sigma$ (lo cual, aunque no se haya caído en la cuenta, es muy razonable, puesto que la raíz cuadrada no es un operador lineal). En realidad se tiene que $E(s) = c_4\sigma$, siendo c_4 una constante cuyo valor depende de los grados de libertad con que se ha estimado s. Para 9 y 11 grados de libertad (los usados para estimar s-within y s-overall) los valores son[2] 0,9727 y 0,9776 respectivamente, por lo que Minitab da, por defecto, los valores: StDev(Within) = 1,62558 y StDev(Overall) = 4,32906. Como es posible quitar esta opción de aplicar la constante para insesgamiento, nosotros la hemos quitado para que los números cuadren con nuestros cálculos.

[2] Más información sobre la constante c_4 y la tabla con sus valores puede encontrarse, por ejemplo, en *Statisitcs for Engineering Problem Solving* de S. B. Vanderman. PWS, Boston 1994.

50 ¿Por qué en los gráficos de control es más eficiente controlar medias que observaciones individuales?

Supongamos que estamos a cargo de un proceso de envasado de paquetes de azúcar con un peso nominal de 1 kg, y para asegurarnos de que el proceso permanece centrado (llena los paquetes con un peso medio igual al valor objetivo) cada 2 minutos pesamos un paquete. Posible pregunta: ¿qué diferencia respecto al valor objetivo nos debe hacer pensar que el proceso se ha descentrado?

Posibles respuestas: 1) Más/menos 10 g, porque esta ya es una diferencia apreciable. 2) Cuando el peso obtenido está fuera de tolerancias, porque es lo que realmente importa.

Las dos respuestas son incorrectas. La primera porque ese valor de 10 g es arbitrario, ¿por qué 10 y no 5, o 15? La segunda porque quizá es posible detectar descentramientos del proceso mucho antes de tener paquetes fuera de tolerancias (esto es lo deseable), aunque también podría ocurrir que la variabilidad de nuestro proceso sea excesiva para poder cumplir con las tolerancias, y en ese caso, por mucho que nos esforcemos en estar continuamente centrando el proceso, no conseguiremos disminuir la variabilidad sino que, al contrario, la aumentaremos[1].

Lo correcto sería hacer un estudio para conocer cuál es la variabilidad intrínseca del proceso (lo que llamamos un estudio de capacidad) y en función de esa variabilidad establecer los límites de control. Por ejemplo, supongamos que de nuestro estudio se deduce que la desviación tipo del proceso (σ) es igual a 5 g. En este caso sabemos que si el proceso está centrado, el 99,7% de los paquetes tendrá un peso situado en el intervalo $1.000 \pm 3\sigma$, es decir, en el intervalo situado entre 985 y 1.015 g.

Por tanto, una buena estrategia será establecer los límites de control en estos valores, y si se obtiene un paquete fuera de límites sospecharemos que el proceso se ha descentrado, sabiendo que el riesgo de tener falsas alarmas (señal de que el proceso está descentrado cuando en realidad no lo está) será del orden del 3 por mil.

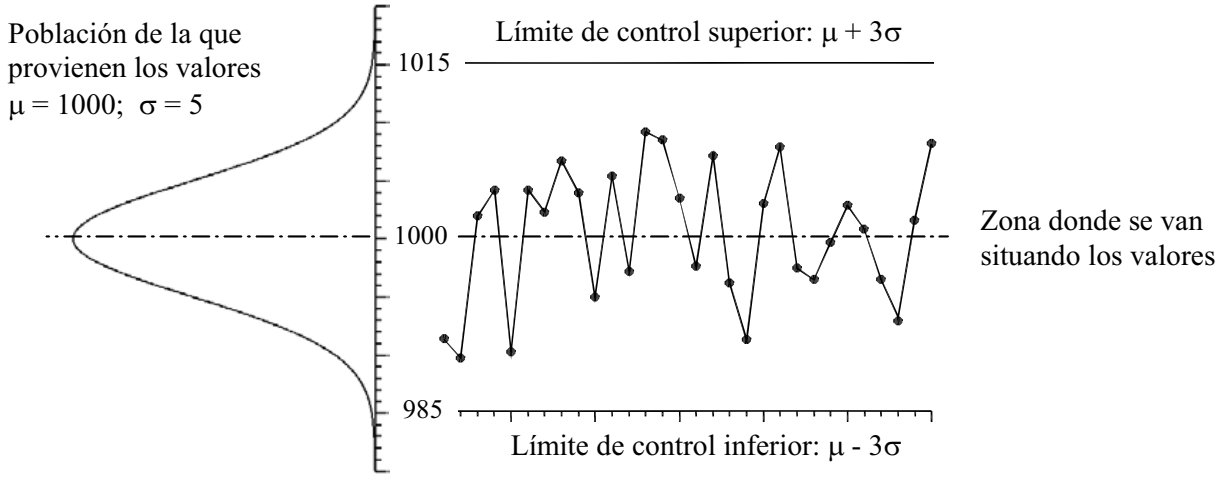


Figura 50.1. *Esquema de sistema de control con límites a ±3σ del valor central*

[1] Es lo que llamamos aumento de variabilidad por sobreajuste. Si vamos moviendo la campana de un lado a otro pretendiendo corregir presuntos descentramientos, lo único que conseguimos es hacer mayor la dispersión de los datos.

Veamos ahora qué ocurre si el proceso se descentra y pasa a envasar los paquetes en torno a 1.010 g. La probabilidad de identificar el descentramiento en la primera medición es igual a la probabilidad de obtener un valor mayor de 1.015 (fuera de límites)[2] de una distribución N(1.010; 5) y esta probabilidad se puede calcular de la forma:

$$z = \frac{x-\mu}{\sigma} = \frac{1.015-1.010}{5} = 1; \qquad P(z>1) = 0{,}16$$

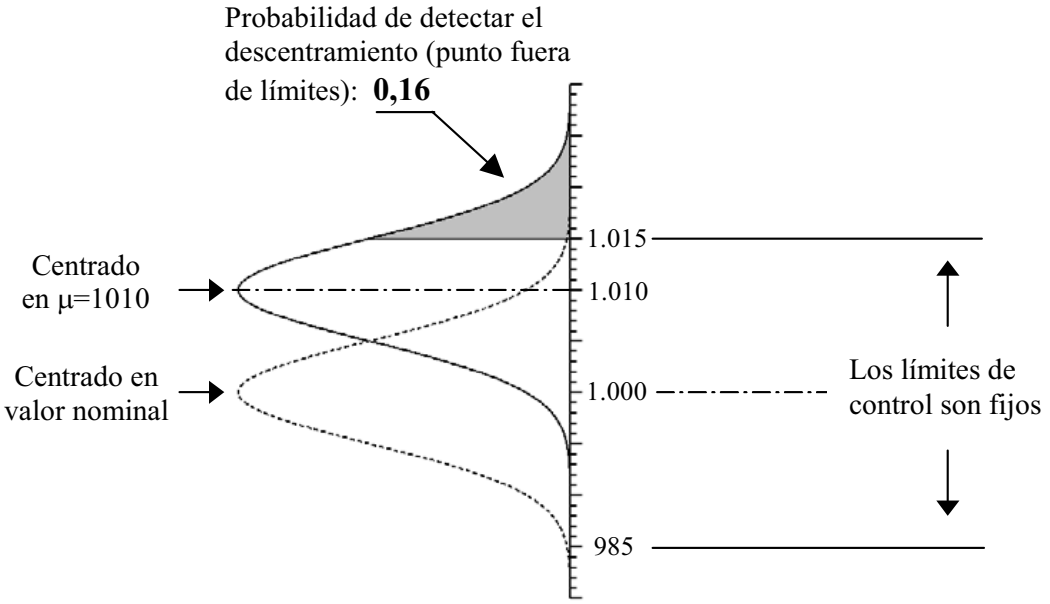


Figura 50.2. *Probabilidad de detectar el descentramiento controlando unidades individuales*

Por tanto, la probabilidad de detectar el descentramiento al hacer la primera medición es 0,16. Veamos ahora cuál es la probabilidad de detectarla dentro del periodo en que se realizan las 4 primeras mediciones después del descentramiento. Podemos hacer:

Probabilidad de detectar el descentramiento = 1 – Probabilidad de no detectar

Probabilidad de no detectar = $(1\text{-}0{,}16)^4 = 0{,}50$

Luego la probabilidad de detectar el descentramiento en ese periodo en que se toman las 4 primeras mediciones es 1-0,50 = 0,50.

Veamos ahora cuál es la probabilidad de detectarlo en la primera medición si lo que se controla son los pesos medios de 4 paquetes consecutivos. En este caso, los límites de control estarán adaptados a la distribución de las medias, y como sabemos que $\sigma_{\bar{X}} = \sigma/\sqrt{n}$, en nuestro caso resultará que la desviación tipo de las medias es $\sigma_{\bar{X}} = 5/\sqrt{4} = 2{,}5$, y por tanto los límites deberán estar a 1.000 ± 7,5 g.

[2] Se calcula solo la probabilidad de que salga por encima del límite superior, ya que la probabilidad de que salga por debajo del límite inferior es despreciable.

Si el proceso se descentra y pasa a envasar en torno a 1.010 g, la distribución de las medias de muestras de tamaño $n = 4$ será N(1.010; 2,5) y la probabilidad de que una media salga de límites es:

$$z = \frac{x - \mu}{\sigma} = \frac{1.007,5 - 1.010}{2,5} = -1; \qquad P(z > -1) = 0,84$$

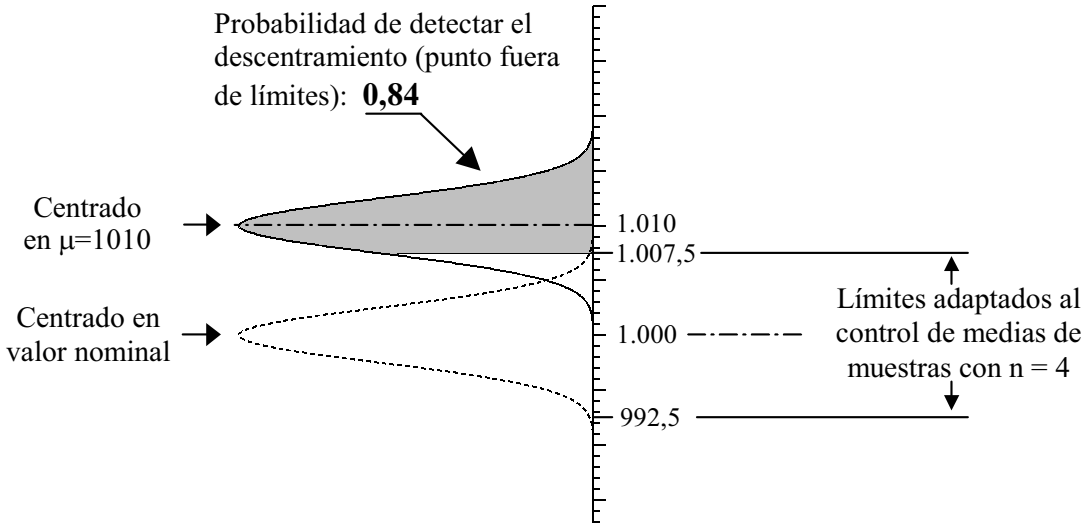


Figura 50.3. *Probabilidad de detectar el descentramiento cuando se controlan medias de 4 observaciones*

Es decir, la probabilidad de detectar que se ha producido el descentramiento después de pesar 4 paquetes es de 0,5 si llevamos sus pesos a un gráfico de valores individuales, y de 0,84 si llevamos la media de esos pesos a un gráfico de medias. Ver el esquema de la Figura 50.4.

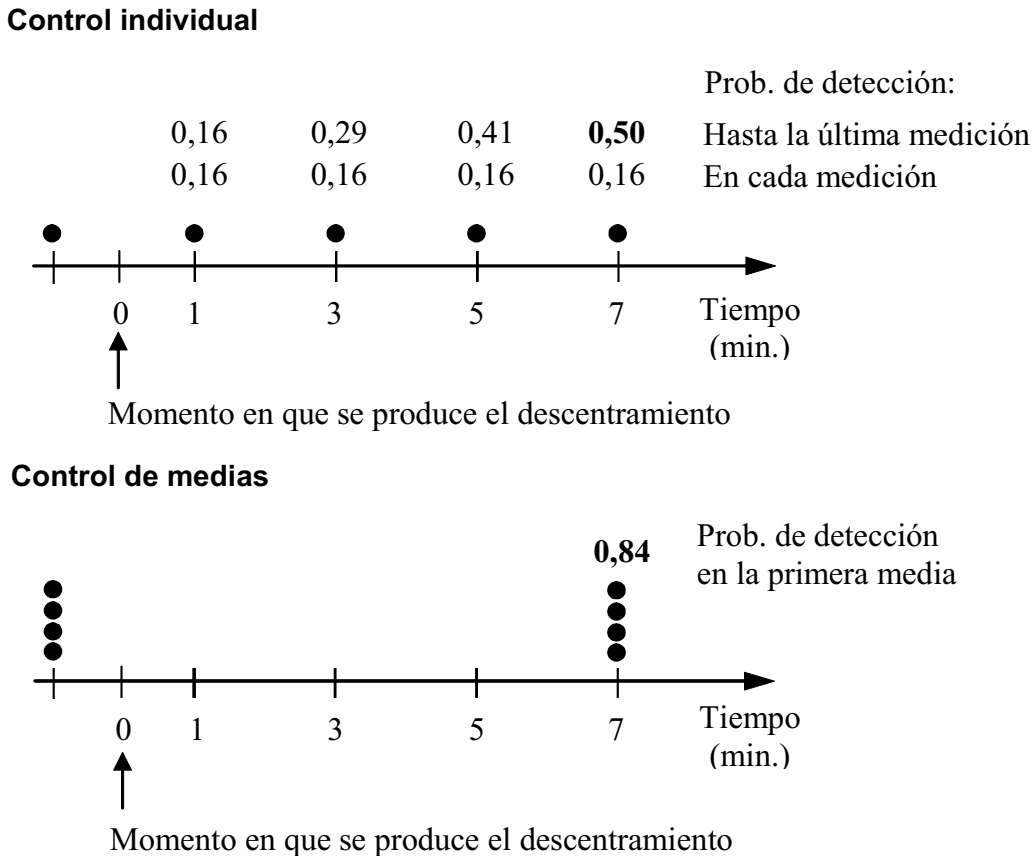


Figura 50.4. *Comparación de las probabilidades de detección controlando medias de 4 observaciones y controlando observaciones individuales*

Podría objetarse que en el caso del control por medias estamos mucho tiempo sin controlar y si, por ejemplo, el descentramiento se produce 1 minuto después del último control, con el sistema de control de medias es imposible enterarse hasta pasados 7 minutos (cuando se produce el próximo control), mientras que con el control individual esta detección ya puede producirse dentro de tan solo 1 minuto.

Efectivamente esta circunstancia se puede dar, pero lo relevante es conocer el tiempo medio que pasará antes de detectar el descentramiento con ambas estrategias. Llamando T_1 al tiempo que pasa desde el descentramiento hasta la detección con el control individual, y $E(T_1)$ a su esperanza matemática, tenemos que:

$$E(T_1) = 1 \cdot 0{,}16 + 3 \cdot 0{,}16 \cdot 0{,}84 + 5 \cdot 0{,}16 \cdot 0{,}84^2 + ...$$

Escrito de forma compacta y completa:

$$E(T_1) = \sum_{i=1}^{\infty} (2i-1) \cdot 0{,}16 \cdot 0{,}84^{i-1}$$

Y haciendo las operaciones (una hoja de cálculo tipo Excel resulta muy útil en estos casos. No es necesario, naturalmente, llegar al infinito) se obtiene que:

$$E(T_1) = 11{,}50 \text{ minutos}$$

Llamemos ahora T_2 al tiempo que pasa desde el descentramiento hasta su detección controlando medias de 4 observaciones. Ahora tendremos:

$$E(T_2) = 7 \cdot 0{,}84 + 15 \cdot 0{,}84 \cdot 0{,}16 + 23 \cdot 0{,}84 \cdot 0{,}16^2 + ...$$

Escribiéndolo de forma compacta y calculando se obtiene:

$$E(T_2) = \sum_{i=1}^{\infty} [(7 + 8(i-1)] \cdot 0{,}84 \cdot 0{,}16^{i-1} = 8{,}52 \text{ minutos}$$

Es decir, en el caso de que se descentre 1 minuto después del último control, y a pesar de que el control individual tiene 3 oportunidades de detección antes que el control de medias, en promedio detectaremos antes el descentramiento controlando medias que observaciones individuales. Si el descentramiento se produjera, por ejemplo, en los minutos 2, 4 o 6, el tiempo medio hasta la detección con el control individual seguiría siendo 11,50 minutos, mientras que controlando medias sería de 7,52; 5,52 y 3,52 minutos respectivamente.

Está claro que controlar medias es más eficiente. Además, el hecho de que se tomen muestras de más de una unidad permite calcular no solo su media sino también el rango, mediante el cual se puede controlar si la dispersión del proceso también se mantiene dentro de los límites establecidos para su variación prevista. Por esta razón los gráficos de control más conocidos son los $\bar{X} - R$, es decir, los de medias y rangos.

51 En los gráficos de control, ¿la línea central debe ser el valor objetivo o el promedio obtenido al hacer el estudio de capacidad?

Recordemos que el control estadístico de un proceso se inicia cuando ya se ha conducido al estado de control, es decir, cuando se da por bueno su valor central y su variabilidad.

El objetivo es detectar si la salida del proceso empeora (o mejora, ¡vamos a ser optimistas!), ya sea cambiando su variabilidad o el valor de la media. Por tanto, la referencia debe estar marcada por la situación inicial real, y no por la ideal que se podría tener.

Veamos un ejemplo. Supongamos que el valor nominal es 10, pero se centra en 11 con $\sigma = 1$ y se da por bueno porque es difícil ajustar más. Si se realiza un control de observaciones individuales y se obtiene el valor $x = 13{,}2$, ¿hay que ajustar? Evidentemente no, porque este valor no es síntoma de que el proceso se haya descentrado con los criterios habitualmente utilizados. Sin embargo, si el valor central se hubiera fijado en el nominal estaríamos a más de 3σ y habría que ajustar, lo cual sería una arbitrariedad, ya que no se tiene ninguna evidencia de que haya empeorado.

En definitiva, la línea central debe ser el promedio de valores obtenidos, tal y como explican la mayoría de libros que tratan este tema.

Varios

52 Cuando se habla de transformación logarítmica, ¿se refiere al logaritmo decimal o al neperiano?

Si el objetivo que se pretende es la normalidad de los datos o la homogeneidad de su varianza, es indiferente cuál sea la base de la transformación logarítmica.

Recordaremos que se dice que *z* es el logaritmo en base *a* de *y*, si y solo si: $a^z = y$, es decir que:

$$Log_a(y) = z \quad \Leftrightarrow \quad a^z = y$$

Para el caso del logaritmo llamado neperiano, la base es el famoso número *e*. Lo denotaremos por *Ln*, en lugar de *Log*. En este tendremos:

$$Ln(y) = x \quad \Leftrightarrow \quad e^x = y$$

Comparemos ahora el *Ln(y),* con el logaritmo en cualquier otra base, por ejemplo en base 10. Tenemos:

$$x = Ln(y) \quad \Leftrightarrow \quad e^x = y$$
$$z = Log_{10}(y) \quad \Leftrightarrow \quad 10^z = y$$

Por tanto, podemos escribir:

$$e^x = 10^z$$

$$Ln(e^x) = Ln(10^z)$$

$$x = Ln(10^z) = z \cdot Ln(10)$$

Como *x* es el nombre que hemos dado a *Ln(y)*, el *Ln*(10) es una constante igual a 2,3026 y *z* es el nombre asignado a $Log_{10}(y)$, tenemos:

$$Ln(y) = z \cdot Ln(10) = 2{,}3026 \cdot z = 2{,}3026 \cdot Log_{10}(y)$$

Es decir, que el logaritmo neperiano se convierte en logaritmo en base 10 al multiplicarlo por una constante, lo cual significa que si al aplicar una transformación con logaritmo neperiano, se logra la distribución Normal o la homogeneidad de varianza, idéntico efecto surte la aplicación del logaritmo en cualquier otra base, en particular en base 10.

53 ¿Qué significan los llamados "grados de libertad"?

Aunque el concepto de grados de libertad está muy presente en los métodos estadísticos, los libros de texto no suelen entrar en él con mucho detalle, quizá porque hacerlo implica meterse en el terreno de la geometría vectorial y esto complica las cosas, además de apartarse de los objetivos que se pretenden en un curso de Estadística. Nosotros, siguiendo con la intuición como principal arma, vamos a presentarlo primero desde un punto de vista geométrico, para después pasar al contexto estadístico.

Visión geométrica

Si le piden que elija un par de números (x, y) al azar, usted tiene libertad completa para la elección de estos dos números. Las dos coordenadas pueden ser representadas por un punto localizado en el plano XY, el cual es un espacio bidimensional. El punto es libre de moverse en ambas direcciones, tiene dos grados de libertad.

Ahora supongamos que nos ponen a elegir un par de números cuya suma es 7. Está claro que solo puede elegirse libremente un número, pues el segundo queda fijado una vez se conozca el primero. Aunque aquí también hay dos variables, solo una es independiente, por lo que el número de grados de libertad se reduce de dos a uno. El punto ahora es libre de moverse en el plano XY pero restringido a permanecer sobre la recta $x+y=7$. Esta línea es un espacio unidimensional que está contenido en el espacio bidimensional original.

Supongamos ahora que nos piden escoger un par de números, tal que la suma de sus cuadrados sea 25. De nuevo, está claro que solo somos libres de escoger uno de los números, pues una vez seleccionemos el primero el otro queda fijado. El punto en cuestión permanece en una circunferencia de radio 5 y con centro en el origen. La circunferencia es un espacio unidimensional contenido en un plano bidimensional. El punto solo puede moverse hacia delante o hacia atrás a lo largo de la circunferencia y por eso tiene un solo grado de libertad. Hay dos números escogidos (N = 2) sujetos a una restricción (r = 1) y el número resultante de grados de libertad es $N-r=2-1=1$.

Supongamos ahora que imponemos simultáneamente las dos condiciones $x+y=7$ y también $x^2+y^2=25$. Si nosotros resolvemos algebraicamente estas ecuaciones, obtenemos que solo son posibles dos soluciones $x=3$, $y=4$ o bien $x=4$, $y=3$. Ninguna variable puede escogerse a voluntad. El punto está restringido por la ecuación $x+y=7$ a moverse a lo largo de una recta, y además está restringido por la ecuación $x^2+y^2=25$, a moverse a lo largo de una circunferencia. Las dos restricciones simultaneas lo confinan a la intersección entre la recta y la circunferencia, dejándolo sin libertad de movimiento, es decir, sin grados de libertad. Aquí, son dos los números a elegir (N=2) y dos las restricciones impuestas (r=2). El número de grados de libertad es $N-r=2-2=0$.

Estas ideas pueden ser generalizadas para N más grande que 2, de forma que cualquier conjunto de N números determinan un punto en el espacio N-dimensional. Si no se

impone ninguna restricción, cada número es libre de variar independientemente de los otros, y por lo tanto el número de grados de libertad es *N*. Cada relación necesaria impuesta sobre ellos reduce el número de grados de libertad en 1. Cualquier ecuación de primer grado que conecte las *N* variables es un espacio de *N*-1 dimensiones. Si por ejemplo, consideramos solamente los puntos cuya suma de coordenadas es una constante, $\sum x_i = c$, hemos limitado el punto a un espacio de *N*-1 dimensiones. Si consideramos solo los puntos que cumplan con $\sum \left(x_i - M\right)^2 = k$, que corresponde a la superficie de una "hiperesfera" con centro en el origen y radio $\sqrt{k}$, esta superficie es un espacio de dimensión *N*–1 dentro de un espacio original de dimensión *N*, y por lo tanto el número de grados de libertad debería ser *N*–1.

También una muestra de tamaño *N* puede ser representada por un punto $\left(X_1, X_2, X_3, \ldots X_N\right)$ en un espacio *N*-dimensional con *N* grados de libertad, cuando no se impone ninguna restricción a sus coordenadas. Ahora bien, si forzamos que su media debe ser $\overline{X}$, ya tenemos una restricción y, por tanto, *N*-1 grados de libertad. Todas las muestras de tamaño *N* con la misma media $\overline{X}$, estarán representadas por los puntos que pertenecen al "hiperplano", $\sum X_i = N\overline{X}$, que será un espacio de (*N*-1) dimensiones.

Grados de libertad en la estimación de la varianza poblacional

Seguramente las primeras reflexiones sobre la idea de grados de libertad en el contexto de la estadística surgen cuando nos intentamos explicar por qué para calcular la varianza muestral se recomienda dividir por *n*-1, en lugar de dividir por *n*, siendo este el número total de datos que compone la muestra. $S^2 = \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 \Big/ (n-1)$

Pero sabemos que la suma de las desviaciones de los datos con respecto a su media es siempre nula, es decir que: $\sum_{i=1}^{n} \left(x_i - \overline{x}\right) = 0$. Esto significa que de los *n* sumandos que tiene el numerador de la varianza solo (*n*-1) son independientes, es decir, solo podemos "inventarnos" *n*-1 si queremos que los *n* tengan la media definida $\overline{x}$.

Por esta razón, aunque la varianza se ha calculado a partir de *n* datos, es importante el hecho de que estos tengan solo *n*-1 grados de libertad, ya que para obtener un estimador insesgado de la varianza poblacional, debemos dividir la suma de cuadrados por el número de sus grados de libertad, y no por el número de datos.

Grados de libertad en la modelización de ecuaciones de regresión

Veamos ahora qué ocurre en el caso de querer ajustar una línea recta a un conjunto de puntos usando el método de los mínimos cuadrados, que consiste en elegir de todas las posibles rectas aquella que haga menor la suma de cuadrados de los residuos e_i (distancia de los puntos a la recta ajustada).

La familia de modelos a considerar es de la forma: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ y de entre todas las posibles rectas, queremos aquella con coeficientes b_0, b_1 que haga mínima la suma

de los cuadrados de los residuos *SCR*. Para esto se realiza un proceso de optimización matemática, obteniéndose que dichos valores b_0, b_1 deben cumplir con las siguientes restricciones:

$$e_1 + e_2 + e_3 + ...e_{n-1} + e_n = 0$$

$$e_1x_1 + e_2x_2 + e_3x_3 + ...e_{n-1}x_{n-1} + e_nx_n = 0$$

La primera restricción quita un grado de libertad a los residuos, ya que conocidos cualesquiera *n*-1 queda unívocamente definido el enésimo. Además, esto implica que la recta debe pasar por el punto $(\overline{x}, \overline{y})$.

Podemos imaginarnos que las posibilidades se restringen a un haz de rectas que pasan por $(\overline{x}, \overline{y})$, pero de todas ellas escogeremos aquella que cumpla con la segunda restricción, con lo cual el error pierde otro grado de libertad, quedando con *n*-2.

Es conveniente saber que $SCR/(n-2)$ es un estimador insesgado de la varianza del error σ^2 y que, en general, para un modelo de regresión múltiple con *p* parámetros (*p*-1 variables predictoras), el número de grados de libertad de los residuos es (*n-p*) y por lo tanto $SCR/(n-p)$ es también un estimador insesgado para la varianza del error.

Generalización del concepto de grados de libertad

Una definición formal de grados de libertad, muy orientada a su cálculo podría ser: "Número de unidades independientes de información en una muestra, que son relevantes para la estimación de un parámetro o para el cálculo de un estadístico."

En el ejemplo del estadístico 'varianza' existe una sola restricción, por lo que los grados de libertad se reducen a *n*-1, y el número de unidades de información que consideramos a efectos de la estimación de la varianza poblacional son *n*-1. De la misma forma, en un modelo de regresión con *n* puntos y *p* parámetros los residuos tienen *n-p* grados de libertad, por lo que en la estimación de la varianza de los errores σ^2 tenemos *n-p* unidades independientes de información y dividimos la suma de cuadrados por *n-p*.

El número de grados de libertad también forma parte de la descripción de algunas distribuciones de probabilidad, que toman esta característica de su relación con ciertos estadísticos. Así, por ejemplo, si $X \sim N(\mu, \sigma)$ y *S* es un estimador de σ calculado a partir de una muestra con *n*-1 grados de libertad, entonces la variable $(X-\mu)/S$ se distribuye según una *t* de Student con *n*-1 grados de libertad (los de *S*), y para describir la variabilidad que presenta la varianza muestral S^2 utilizamos la distribución Chi-cuadrado, también con *n-1* grados de libertad (los de S^2). Finalmente, si tenemos 2 muestras de tamaños n_1 y n_2 de una población Normal, el cociente S_1^2 / S_2^2 sigue una distribución F de Snedecor con n_1-1 (los de la varianza del numerador) y n_2-1 (los de la varianza del denominador) grados de libertad.

54 ¿Debe decirse “Teorema central del límite” o “Teorema del límite central”?

Lo central, en el sentido de clave o fundamental, es el teorema, no el límite, que es como todos los límites. Debe decirse, por tanto, “Teorema central del límite” y no de la otra forma.

¿Y por qué este teorema es central en la teoría estadística?

Cuando se tiene la suerte de enfrentar un fenómeno cuya característica de interés se puede modelar con una distribución Normal, heredamos de inmediato un arsenal de resultados útiles para la estimación de los parámetros y para hacer contraste de hipótesis. Si la aplicación de toda esa valiosa teoría dependiera de la suerte de toparse con variables de distribución Normal, su aplicación quedaría bastante restringida en la solución de problemas reales, pues existe también un amplio abanico de fenómenos en la naturaleza que no se ajustan a la distribución Normal.

Lo que le da un potente valor agregado a la batería de resultados de la inferencia estadística basada en la ley Normal, es saber que gran parte de dicho conocimiento es útil, aun cuando la característica en estudio no se pueda modelar con esta distribución. Esto se justifica precisamente con el teorema a que estamos haciendo referencia, pues demuestra que es plausible modelar la variable aleatoria media muestral por medio de la distribución Normal, aun cuando la población madre de donde se extrae la muestra tenga otra distribución distinta de la Normal, exigiendo para su aplicación un conjunto no muy fuerte de condiciones que se suelen cumplir en la práctica.

Hay pocos resultados tan importantes como este, por eso se trata de un teorema central en la teoría estadística.

55 ¿Cuál es la mejor estrategia para ganar la lotería (nacional, primitiva,...)?

Si nos referimos solo a juegos de azar que se basan en la extracción de bolas de un bombo, esta pregunta no tiene respuesta, ya que no existe una estrategia mejor que otra. Y es una lástima, porque a muchos les habrá parecido que esta era la pregunta más interesante del libro.

Pero por no hacer esta respuesta demasiado corta, haremos algunas consideraciones sobre los sorteos en los que se compra un boleto con un número (lotería nacional, sorteos de la ONCE,…), y aquellos en los que se elige una combinación de 6 entre los números 1 al 49 (lotto, lotería primitiva,…).

Boleto de lotería con un número

Supongamos que los números disponibles son del 0 al 99.999 y que se elige uno al azar. Es evidente que si el número se saca al azar, todos tienen la misma probabilidad de salir, y aquí no hay táctica que valga. Sin embargo, existen algunas falacias bastante extendidas que podemos comentar. Por ejemplo:

- *Si se compra en determinados lugares (ciudades, puntos de venta) es más fácil que toque que si se compra en otros.* Los que mantienen esta postura la avalan con datos: en ese lugar que ellos dicen que da más premios, efectivamente en los últimos años ha dado muchos más que en ese otro donde es muy difícil que toque. La clave está en distinguir el número de premios que da, de la probabilidad de que toque. El número de premios que caen en un lugar depende del número de boletos que se han vendido. Por eso, a la larga, tocan más premios en Madrid o Barcelona que en Teruel, pero esto no implica que tengamos más posibilidades de que nos toque si compramos el número en Madrid o Barcelona. Las probabilidades son exactamente las mismas. Respecto a los puntos de venta, si en uno de ellos toca algún premio importante y se crea el clima de opinión de que allí es fácil que toque, la gente irá a comprar más allí y, al venderse más números, efectivamente aumenta la probabilidad de que allí vuelva a tocar, aunque para el que compra la probabilidad no cambia en absoluto, tanto si lo compra allí como en cualquier otro lugar.

- *Hay números "feos" que no hay que comprar porque nunca tocan*. Si a uno le ofrecen el número 01.010 es normal que no lo quiera porque este es un número raro y le parece muy difícil que toque. Preferirá el 34.278 que es mucho más normal. Después resulta que toca un número como el 47.121, lo cual confirma la teoría de que los números que salen son normales, aunque, lamentablemente, no el "normal" que habíamos comprado. ¿Por qué toca muy pocas veces a los números raros? Pues simplemente porque números raros hay menos. Visto de otra forma: ¿Por qué es más probable que toque fuera del intervalo 30.000–35.000 que dentro? Pues simplemente porque fuera hay más números que dentro, pero esto no hace que

menospreciemos los números dentro de este intervalo. Lo mismo ocurre con los números raros[1].

Consejos: Si compra lotería, hágalo cerca de su casa o donde le pille más a mano. Se ahorrará tiempo, gastos de desplazamiento, y quizá el disgusto de comprobar que ha tocado en el punto de venta de su calle y usted se la fue a comprar a otra ciudad. Y no rechace los números raros, especialmente si los ha visto y se va a acordar del número.

Elegir 6 números del 1 al 49

Este es un juego muy popular en todo el mundo. Se conoce con muy diversos nombres como Lotería primitiva, Lotto, Loto 6/49, …. Hay premio a partir de 3 aciertos. Las probabilidades de acertar son[2]:

Número de aciertos	Probabilidad
3	1 en 57
4	1 en 1032
5	1 en 55.491
5 + Complementario	1 en 2.330.636
6	1 en 13.983.816

Algunas creencias sin fundamento son:

- *Hay que apostar a los números que han salido menos, ya que como la probabilidad es la misma para todos y a la larga su frecuencia de aparición tiende a igualarse, será más probable que salgan los que menos han salido.* Es verdad que a la larga

[1] Por ejemplo, si por raro entendemos un número con solo 1 o 2 cifras distintas (por ejemplo, el 22.222 o el 22.552 respectivamente), números en los que aparece solo una cifra tenemos 10 (00.000, 11.111, … 99.999) y números en los que aparecen 2 cifras, si tenemos 4 iguales y una distinta, son: $10 \cdot 9 \cdot 5 = 450$, (10 posibilidades para el primero, 9 para el segundo y 5 formas posibles de ordenarlos). Si tenemos 3 iguales por un lado y 2 iguales por otro: $10 \cdot 9 \cdot (5! / 2! \cdot 3!) = 900$. Es decir, que con este criterio, números raros hay 10 + 450 + 900 = 1.360 entre 100.000.

[2] Para calcular estas probabilidades podemos usar la regla de casos favorables partido por casos posibles. Los casos posibles, es decir, las formas de elegir 6 números de un conjunto de 49 son las combinaciones de 49 tomados de 6 en 6, que notamos $\binom{49}{6}$. Los casos favorables para acertar 3 es el producto de los casos en que se pueden elegir los 3 que se aciertan de entre los 6 premiados $\binom{6}{3}$ por los casos en que se pueden elegir los 3 que no se aciertan de entre los 43 no premiados $\binom{43}{3}$. El resultado para la probabilidad de acertar 3 es: $\binom{6}{3} \cdot \binom{43}{3} \Big/ \binom{49}{6} =$ 0,0176504. Análogamente, para 4 aciertos la probabilidad es $\binom{6}{4} \cdot \binom{43}{2} \Big/ \binom{49}{6} =$ 0,0009686. La probabilidad de acertar 5, sin el número complementario, es $\binom{6}{5} \cdot \binom{42}{1} \Big/ \binom{49}{6} =$ 0,0000180 (en el numerador aparece 42 porque el número no premiado se elige de entre los no premiados menos el complementario). La probabilidad de acertar 5 más el complementario es $\binom{6}{5} \Big/ \binom{49}{6} =$ 0,0000004, ya que hay una sola forma de elegir el complementario. Y finalmente, la probabilidad de acertar los 6 es una entre todas las combinaciones posibles, es decir: $1 \Big/ \binom{49}{6} =$ 1 entre 14 millones, aproximadamente.

las frecuencias de aparición tienden a igualarse, de la misma forma que si lanzamos muchas veces una moneda al aire, cuantas más veces la lancemos más cerca estará del 50% la proporción de caras y cruces. Pero esto no quiere decir que después de 5 caras sea más probable que salga una cruz, la probabilidad siempre es constante para ambos resultados y de la misma forma ocurre con los números que salen en este tipo de sorteos. Los bombos no tienen memoria, y no saben lo que ha salido antes. Ni memoria ni sentimientos de justicia o igualdad. Simplemente las probabilidades son las mismas en todos los sorteos, independientemente de lo que haya salido en los sorteos anteriores.

- *Hay que apostar por los números que más salen, puesto que esos, por las razones que sean, son los que han demostrado tener más probabilidad de salir*. Lo verdaderamente sorprendente sería que todos los números hubieran salido exactamente el mismo número de veces. Lo normal es que, por azar, unos hayan salido más y otros menos. Pero no hay ninguna razón para pensar que los que han salido más hasta ahora vayan a seguir saliendo más. Por la misma razón que antes las probabilidades son idénticas en cada sorteo.

- *Existen estrategias que con ayuda de un ordenador permiten elaborar combinaciones múltiples que aumentan las probabilidades de ganar*. Si por ganar se entiende que tocan premios, cuanto más se juega más probabilidad se tiene de que toquen. Pero si por ganar se entiende tener beneficios (premio-coste del juego), la tendencia es que cuanto más se juega más se pierde, con independencia de la estrategia que se siga. En otro tipo de apuestas, en las que el resultado no depende solo del azar, como las quinielas de fútbol, sí pueden existir estrategias para aumentar la probabilidad de beneficio.

Consejos: Aunque la probabilidad de que una combinación toque es siempre la misma, cuanto mayor sea la cantidad a repartir más dinero ganará si le toca, por lo que tiene más emoción jugar cuando hay mucho dinero acumulado de otras semanas en las que no hubo ningún acertante. Por otra parte, como el premio se reparte entre los acertantes, parece una buena idea apostar por una combinación rara, del tipo 1, 2, 3, 4, 5 y 6, ya que tiene las mismas probabilidades de salir que cualquier otra, y si nos toca probablemente seremos los únicos acertantes y cobraremos más dinero. Un estudio realizado sobre la Lotto Canadiense puso de manifiesto que la apuesta menos popular consistía en los números: 20, 30, 39, 40, 41 y 48, y la más popular (la que sería peor apuesta) era: 3, 7, 9, 11, 25 y 27. Aunque quizá esto era cierto antes de que se hiciera público pero no ahora. Además, seguramente el que haya más o menos tendencia a marcar un número también depende de cuál es la distribución de los números en el boleto y no estamos seguros de que sea la misma en todas partes. En definitiva, marque los números que se le antojen o, todavía más cómodo, que los marque la máquina, y espere a ver si ha habido suerte, con la seguridad de que nadie ha usado una estrategia mejor que la suya[3].

[3] Un libro interesante, que comenta el estudio realizado sobre la Lotto Canadiense y que explica las probabilidades de ganar en diferentes juegos de azar es *Can You Win?* de Mikel Orkin. El autor demuestra un gran dominio en estos temas, y la contraportada del libro indica que es Director del Departamento de Estadística de la "California State University", y un conocido experto en cálculo de probabilidades y estadística que aparece regularmente en radio y televisión. Pero no se sabe si estos conocimientos le han servido para ganar alguna vez la lotería.

Créditos y referencias

¿Cómo hemos resuelto nuestras dudas?

La respuesta a las preguntas que planteamos no siempre ha sido tan fácil como habíamos previsto al principio. En algunas nos ha costado encontrar un equilibrio que nos gustara entre rigor, sencillez e intuición. En otras previamente hemos tenido que aclarar y consensuar nuestras ideas, ya que incluso la respuesta a preguntas que parecen de principiantes pueden despertar controversias.

En el proceso de resolver nuestras dudas y de buscar la mejor manera de dar las explicaciones, hemos contado con la ayuda inestimable de nuestros compañeros, y también con la de libros, artículos y páginas web, que tratan los temas que nosotros planteamos. A continuación citamos las fuentes de información que hemos utilizado o que pensamos han tenido influencia en nuestras respuestas. Lo hacemos con una doble intención: la de que quede la constancia y nuestro reconocimiento a las que han sido nuestras fuentes, y también para que el lector interesado pueda acudir a ellas si desea recabar más información.

Estadística descriptiva

Una descripción sencilla y clara de la utilidad de la media geométrica y de la media armónica la encontramos en un texto de H. L. Alder y E. B. Roessler: *Introduction to Probability and Statistics,* Freeman 1968. Se trata de un libro de introducción general a las técnicas estadísticas que a pesar de los años pasados desde que fue escrito sigue valiendo la pena tenerlo presente.

Los ejemplos para mostrar que dividiendo la varianza muestral por *n* se obtiene un estimador sesgado de la varianza poblacional los vimos planteados en el texto de Alan Stuart *The Ideas of Sampling Monograph Series No. 4;* Charles Griffin, 1984.

Los histogramas constituyen una de las llamadas *7 Herramientas básicas de Ishikawa,* y algunos aspectos de nuestra respuesta sobre como construirlos son los que propone este autor en su libro *Guía de Control de Calidad,* UNIPUB, 1985, uno de los clásicos sobre control y mejora de la calidad.

Para responder a las preguntas sobre cómo calcular los cuartiles y porqué se utiliza 1,5 veces el rango intercuartílico para definir la zona de anomalías en los boxplots, ha sido una excelente fuente de información la página web *http://exploringdata.cqu.edu.au* de la Central Queensland University (Australia) preparada por Rex Boggs. Hay mucho y excelente material en este sitio para explicar análisis exploratorio de datos en cursos introductorios de estadística.

Los datos sobre la velocidad de paso de los coches en función del tipo de señal disuasoria, está inspirado en unos que incluye Minitab (*Student1*) aunque las anomalías las hemos puesto nosotros.

Un buen tratamiento sobre los coeficientes de apuntamiento (*kurtosis)* y de asimetría (*skewness)* describiendo con detalle las limitaciones de los estadísticos en la estimación

de sus respectivos parámetros, y que nos ha sido útil para redactar nuestra respuesta, se encuentra en el texto de D.J. Wheeler *Advanced Topics in Statistical Process Control,* SPC Press 1995.

Distribuciones de probabilidad

Que la distribución del coeficiente de correlación para muestras de tamaño 3 tiene forma de U lo descubrimos en el excelente libro de Alder y Roessler antes citado.

La deducción de la fórmula de la distribución de Poisson a partir de la binomial está en muchos libros, pero para explicar todos los pasos con detalle nos ha sido útil el de K. Knopp. *Theory and Application of Infinite Series,* Dover, 1990.

Para justificar que la distribución de la varianza muestral está relacionada con la distribución Chi cuadrado hemos seguido el esquema que presentan R. E. Walpole y R. H. Myers en su libro *Probabilidad y Estadística* McGraw-Hill, 1991.

Estimación

Un trabajo excelente, que ha sido fuente de inspiración para nuestra respuesta al porqué cuesta acertar en los sondeos electorales, es el artículo de F. Udina y P. Delicado: ¿Cómo y cuánto fallan los sondeos electorales?, publicado en la *Revista Española de Investigaciones Sociológica, núm. 96,* 2001.

Contraste de hipótesis

En las preguntas sobre contraste de hipótesis hemos sido influenciados por el excelente libro de Angustias Vallecillos *Inferencia Estadística y Enseñanza: Un análisis didáctico del contraste de hipótesis estadísticas*. Editorial Comares (1996), en el cual se analizan los aspectos conceptuales y filosóficos del contraste de hipótesis y se presentan los resultados de una seria investigación empírica sobre lo que comprenden los estudiantes en este tema.

Comparación de tratamientos

La tabla que se incluye en la respuesta a la pregunta sobre ¿cómo se sabe hacia que lado hay que mirar el área de la cola? está tomada del libro de A. Prat *et al. Métodos estadísticos. Control y mejora de la calidad.* Ediciones UPC, 1997. Editado en Latinoamérica por Alfaomega.

Correlación y Regresión

Los valores que aparecen en los gráficos de las Figuras 35.1 y 35.2 para mostrar que es más adecuado minimizar la suma de los cuadrados de los residuos que su valor absoluto están en el texto de T.H. Wonnacott y R.J. Wonnacott: *Introducción a la estadística.* Limusa, 1979. Nos ha parecido que estos datos combinan muy bien la sencillez con la claridad.

La idea de superponer las rectas obtenidas con varios conjuntos de datos para mostrar la variabilidad que presentan los coeficientes (Figuras 36.1 a 36.3) se nos ocurrió a la vista de una de las *applets* desarrolladas en el proyecto "*VESTAC: Visualization of and Experimentation with STAtistical Concepts"* desarrollado en la Katholieke Universiteit Leuveny (Universidad Católica de Lovaina, Bélgica) y que se encuentran en *http://www.kuleuven.ac.be/ucs/java/index.htm*. Vale la pena entrar en este sitio y explorar los *applets* que incluye. El que comentamos está en el apartado de regresión y se llama *Histograms of slope and intercept*. El material es excelente y es una suerte que sea de acceso libre en la red.

A nuestra amiga, la profesora Lourdes Pozueta, le oímos explicar la analogía de llevarse asesores a un examen para ilustrar la estrategia de la regresión paso a paso. El ejemplo nos parece muy bueno y nosotros también lo usamos ahora en nuestras clases.

Diseño de experimentos

Existen bastantes libros en los que se presentan los diseños factoriales y sus posibilidades de uso. A nosotros nos gustan el de Box, Hunter y Hunter: *Estadística para Investigadores* traducido al castellano en Editorial Reverté, 1988 y el de Prat *et al.* antes citado.

A Albert Prat y Xavier Tort-Martorell les debemos muchas ideas sobre la forma de explicar diseño de experimentos. El ejemplo que hace referencia a lo crujientes que resultan las galletas está inspirado en un trabajo que realizó Xavier Tort-Martorell.

Para explicar el porqué funciona el algoritmo de Yates teníamos un explicación que después sustituimos por la que aquí aparece, que resulta más clara y más corta. Esta la vimos en el artículo de J. C. Gower: The Yates Algorithm, publicado en la revista *Utilitas Matemática*, *Volumen 21B,* 1982. Arturo De Zan nos ayudó en la búsqueda de material sobre este tema.

Estudios de capacidad y control estadístico de procesos

Daniel Peña y Albert Prat, escribieron en 1986 un libro introductorio sobre la calidad en la empresa con el título: *Cómo mejorar la calidad*, Editado por el Instituto de la Pequeña y Mediana Empresa Industrial. En este texto se muestra la ventaja de controlar medias frente a observaciones individuales utilizando el ejemplo de muestras de tamaño 4 cuando el proceso se descentra 2 desviaciones tipo. Como estos valores son sencillos y resultan muy didácticos, nosotros hemos usado los mismos.

Varios

Para responder a la pregunta sobre lo que son los grados de libertad nos ha sido útil el artículo de Walter. M. Helen: *Degree Freedom* publicado en el *Journal of Educational Psychology, núm. 31, vol. 14* (1940).

Libros y páginas web que se citan

El número entre corchetes indica la pregunta con la que están relacionados.

Agencia de Protección Medioambiental de EE UU. Página web: *www.epa.gov/ceampubl/mmedia/metdata.* [16]

Alder HL, y Roessler, EB.: *Introduction to Probability and Statistics,* Freeman, 1968. [2] [15] [15]

Anderson, TW.: *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, 1994. [16]

Barnett V, y Lewis, T.: *Outliers in Statistical Data*. John Wiley & Sons, 1994. [9]

Bernardo JM.: Monitoring the 1982 Spanish Socialist Victory: A Bayesian Analysis. *Journal of the American Statistical Association*. Vol. 79, Núm. 387 (1984). [23]

Bouza Álvarez, F.: Comunicación política: encuestas, agendas y procesos cognitivos electorales. *Revista 'Praxis Sociológica'* 1998 número 3. [23]

Box G, Hunter W, y Hunter J.: *Estadística para Investigadores* Reverté, 1988. [34] [41] [42] [46] [47]

Cansado, E.: *Estadística General*. Centro Internacional de Enseñanza de la Estadística (CIENES). Santiago de Chile, 1967. [3]

DeGroot, MH.: *Probabilidad y Estadística* Addison-Wesley Iberoamericana, 1988. [13]

Draper, NR, y Smith, H.: *Applied Regression Analysis*. J. Wiley & Sons, 1998. [35]

Education Queensland (Departamento de Educación del Estado de Queensland, Australia). Página web con material de ayuda para profesores de estadística. Preparada por Rex Boggs. *http://exploringdata.cqu.edu.au* [5] [8]

García Ferrando, M.: *Socioestadística. Introducción a la Estadística en Sociología* Alianza Editorial, 1988. [21]

Gower, JC.: The Yates Algoritm Revista *Utilitas Matemática*, Volumen 21B, 1982. [43]

Helen, Walter. M.: Degree Freedom. *Journal of Educational Psychology,* núm. 31, vol. 14 (1940). [53]

Ishikawa, K.: *Guía de Control de Calidad* UNIPUB, 1985. [6]

Knopp, K. *Theory and Application of Infinite Series,* Dover, 1990. [16]

Montgomery, DC.: *Design and Analysis of Experiments*. John Wiley & Sons, 1997. [44]

Moore, David y McCabe, George: *Introduction to the Practice of Statistics*. W H Freeman & Co., 1998 (3ª edición). [5]

Organización Mundial de la Salud. Página web: *www3.who.int/whosis/menu.cfm*. [16]

Orkin, Mikel: *Can You Win*?. W.H. Freeman and Company, 1991. [55]

Peña, D. y Prat, A.: *Cómo mejorar la calidad* Instituto de la Pequeña y Mediana Empresa Industrial, 1986. [50]

Peña, Daniel: *Fundamentos de estadística*. Alianza Editorial, 2001. [30]

Prat A., Tort-Martorell, X., Grima, P. y Pozueta, L.: *Métodos estadísticos. Control y mejora de la calidad*. Ediciones UPC (en España) Alfaomega (en Latinoamérica), 1997. [32] [41] [42] [46] [47]

Stuart, Alan: *The Ideas of Sampling Monograph Series No. 4; Charles Griffin, 1984*. [4]

Tukey, John: *Exploratory Data Analisys* Addison-Wesley, 1977. [5]

Udina F, y Delicado P.: ¿Cómo y cuánto fallan los sondeos electorales? *Revista Española de Investigaciones Sociológica*, núm. 96, 2001. [22]

Vallecillos, Angustias: *Inferencia Estadística y Enseñanza: Un análisis didáctico del contraste de hipótesis estadísticas*. Comares, 1996. [24]

Vanderman, SB.: *Statisitcs for Engineering Problem Solving*. PWS, Boston 1994. [49]

VESTAC*: Visualization of and Experimentation with STAtistical Concepts"*. Página web que contiene un conjunto de applets para visualizar conceptos clave de estadística. Universidad Católica de Lovaina, Bélgica. *www.kuleuven.ac.be/ucs/java/index.htm.* [36].

Walpole, RE, y Myers RH.: *Probabilidad y Estadística* McGraw-Hill, 1991. [17]

Wheeler, DJ.: *Advanced Topics in Statistical Process Control,* SPC Press 1995. [10]

Wonnacott, TH, y Wonnacott, RJ.: *Introducción a la estadística*. Limusa, 1979. [35]

WIRIS. Calculadora simbólica accesible a través de internet. Más información en *http://www.wiris.com* [13]

Yates, F.: *The Design and Analysis of Factorial Experiments*. Technical Communication No. 35. Commonwealth Bureau of Soils. Harpenden, Reino Unido, 1937. [43]

NOTAS

NOTAS

NOTAS

NOTAS